ISSN: 1648-6897 / eISSN: 1822-4199

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of ENVIRONMENTAL ENGINEERING & LANDSCAPE MANAGEMENT

2026
Volume 34
Issue 2
Pages 124–138
https://doi.org/10.3846/jeelm.2026.25791

MODELING THE EFFECT OF POLLUTANT GAS ON $PM_{2.5}$ IN CHINA WITH COMPUTATIONAL INTELLIGENCE

Xu WANG[1], Kai ZHANG[2✉], Peishan HAN[3], Xianjun LI[4], Qiong PAN[5, 6✉✉]

[1]*Shool of Software, Shanxi Agricultural University, Taiyuan, China*
[2]*Chongqing Chang'an Industrial Co. Ltd, Chongqing, China*
[3]*Passenger Transport Third Branch, Shenzhen Metro Operation Group Co. Ltd., Shenzhen, China*
[4]*Meteorological Bureau of Yangling, Yangling, Shaanxi, China*
[5]*School of Science, Northwestern A&F University, Yangling, Shaanxi, China*
[6]*School of Telecommunications Engineering, Xidian University, Xi'an, China*

Highlights:

- four computational intelligence methods (GEP, BPNN, SVR, LR) modeled pollutant gas effects on $PM_{2.5}$;
- pollutant impact on $PM_{2.5}$ ranged from –0.7579 to 0.9802;
- CO and PM_{10} were identified as the top contributors to $PM_{2.5}$;
- GEP and LR formulas support further $PM_{2.5}$ analysis and prediction;
- findings offer insights for $PM_{2.5}$ control and forecasting.

Article History:

- received 26 August 2024
- accepted 29 September 2025

Abstract. This study employs computational intelligence techniques – gene expression programming (GEP), back-propagation neural network (BPNN), support vector regression (SVR) and linear regression (LR)–to model the quantitative relationship between pollutant gases (PGs) and $PM_{2.5}$ concentrations using 2021 environmental data from 12 Chinese cities. A comparative analysis was conducted to evaluate model performance using the correlation coefficient (R), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Results showed that the correlation coefficients (R) between predicted and actual $PM_{2.5}$ concentrations ranged from –0.7579 to 0.9802 across all models. SVR and LR demonstrated the most robust performance, achieving high average R values of 0.8656 and 0.8671, respectively. LR also yielded the lowest average RMSE (0.12) and MAE (0.06) across the cities. GEP proved capable of finding highly accurate explicit models, achieving a maximum R of 0.9766. A key finding from the LR models is that CO and PM_{10} consistently had the most significant impact on $PM_{2.5}$ concentrations. Correlation formulas derived from GEP and LR can support further $PM_{2.5}$ analysis. These findings offer insights into $PM_{2.5}$ formation mechanisms and inform pollution control strategies.

Keywords: $PM_{2.5}$, pollutant gas, GEP, BP neural network, SVR, LR.

✉Corresponding author. E-mail: ***zhang356980@163.com***
✉✉Corresponding author. E-mail: ***qiongqpan@163.com***

1. Introduction

Particles called $PM_{2.5}$, whose diameter is less than 2.5 μm, are a type of atmospheric pollution that is detrimental to the respiratory (Dondi et al., 2023; Onaiwu & Eferavware, 2023; Widziewicz-Rzońca et al., 2022), cardiovascular (Münzel et al., 2021; Yuan et al., 2023), digestive (Dorofeyev et al., 2023), blood (Liu et al., 2023), nervous (López-Granero et al., 2023) and reproductive (Zhang et al., 2024) systems of humans. Moreover, $PM_{2.5}$ carries more than 1600 types of hazardous substances, such as heavy metals that can result in cancer and bacteria that could damage skin or have other more dangerous consequences. Furthermore, $PM_{2.5}$ can induce or aggravate pulmonary inflammation and cause cardiovascular diseases, such as myocardial infarction, by increasing blood viscosity that can subsequently cause thrombus and affect fetal development. There is growing evidence (Xu et al., 2023) to suggest that NO_2, SO_2, CO, O_3 and PM_{10} (Chen et al., 2023) are the principal gaseous constituents that can influence the concentrations of $PM_{2.5}$ under certain environmental conditions. Therefore, establishing a quantitative model of

the correlation between these pollutant gases (PGs) and $PM_{2.5}$ is of great importance.

Numerous studies have explored the application of artificial intelligence (AI) and machine learning (ML) techniques for air pollution prediction, offering significant advantages over traditional approaches due to their ability to uncover complex patterns in large datasets (Arabloo et al., 2015; Kumar et al., 2020; Schweidtmann et al., 2021; Samad et al., 2023). For instance, Samad et al. (2023) demonstrated the potential of ML models to replace physical monitoring stations with virtual ones for air pollution prediction. Similarly, Kumar et al. (2020) proposed an ML-based model for $PM_{2.5}$ concentration estimation in Delhi, highlighting the efficacy of regression and time series analysis. While these studies showcase the predictive power of ML, they often focus on a single model or specific pollutant, lacking a broad comparative analysis across diverse AI/ML paradigms and their ability to generate explicit, interpretable relationships. Other research, such as Kokkinos et al. (2021), compared statistical and computational intelligence methods for traffic-induced particulate matter, yet their focus was primarily on prediction accuracy rather than explicit quantitative correlation and formula derivation for different PGs. Drewil and Al-Bahadili explored LSTM deep learning and metaheuristics for air pollution prediction, demonstrating high accuracy but without emphasizing explicit formula generation or the comparative impact of individual pollutants on $PM_{2.5}$ in a multi-city context (Drewil & Al-Bahadili, 2022). Furthermore, while hybrid models integrating mechanistic and data-driven approaches have been systematically reviewed for chemical and energy systems, their application to deriving specific pollutant-$PM_{2.5}$ relationships remains an area for further investigation (Zendehboudi et al., 2018).

While numerous studies have explored the prediction of $PM_{2.5}$ concentrations using various machine learning techniques, our study addresses a distinct research gap by providing a comprehensive comparative analysis of four diverse computational intelligence paradigms–GEP, BPNN, SVR, and LR–applied specifically to quantify the direct influence of PGs across a geographically varied urban landscape in China. The novelty of this work lies in its unique multi-faceted approach: (1) we rigorously compare modeling techniques ranging from the ‹white-box› evolutionary GEP, which generates explicit mathematical formulas, to the ‹black-box› BPNN and established statistical methods SVR and LR; (2) our primary goal is not just prediction but the derivation of interpretable equations that reveal the quantitative impact of individual pollutants; and (3) we utilize a unique dataset covering 12 cities that span the full spectrum of air quality in China, from severely polluted to among the cleanest, providing a robust testbed for model generalization. The explicit identification of key influencing PGs and the provision of actionable formulas represent significant contributions to the field of air quality modeling and management.

The study aims to build a model to describe the correlation between $PM_{2.5}$ and PGs. There are five kinds of PGs specified in the air quality index of China: SO_2, NO_2, CO, O_3, and PM_{10} (as a precursor and co-pollutant). We collected relevant data (January 1, 2021–December 31, 2021) from 12 cities in China. Some of the cities have a serious problem with $PM_{2.5}$, others less of a problem, and a few have some of the best air quality in China. Four types of computational intelligence methods were adopted in this research: gene expression programming (GEP), back-propagation neural network (BPNN), support vector regression (SVR), and linear regression (LR). This selection was deliberate, aiming to encompass a diverse range of modeling paradigms, each with unique strengths relevant to $PM_{2.5}$ modeling. GEP was chosen for its capacity to generate explicit mathematical formulas, offering high interpretability. BPNN, a widely adopted neural network, provides robust non-linear mapping capabilities. SVR is recognized for its strong generalization performance based on statistical learning theory, particularly effective in handling high-dimensional data. Lastly, LR serves as a fundamental statistical benchmark, providing insights into linear relationships and easily interpretable coefficients for pollutant impact assessment. This comprehensive suite allows for a robust comparative analysis of different modeling approaches in addressing the complex $PM_{2.5}$ phenomenon. Outcomes indicated that the mean correlation coefficients of PGs on $PM_{2.5}$ in each city ranged from 0.6524 to 0.8866 (GEP), –0.2455 to 0.6820 (BPNN), 0.5907 to 0.9607 (SVR), and 0.6299 to 0.9382 (LR). Across the studied cities and based on mean correlation coefficients, SVR and LR generally exhibited the highest average performance, suggesting their robustness in capturing the dominant linear and non-linear patterns within the dataset. GEP also demonstrated strong performance with high maximum correlation coefficients, showcasing its capability to find highly fitting explicit relationships, while BPNN showed greater variability across different city datasets. This comparative analysis provides critical insights into the applicability and performance of different AI/ML approaches for $PM_{2.5}$ modeling. Besides, the equations generated by GEP and LR can portray the relationship and evolutionary law between PGs and $PM_{2.5}$. Therefore, these findings can be used to predict concentrations of $PM_{2.5}$. The results returned by LR can also indicate which pollutants influence the concentrations of $PM_{2.5}$ more significantly. Findings showed that different PGs affect $PM_{2.5}$ to varying degrees; especially, CO and PM_{10} were found to contribute most to $PM_{2.5}$. Ultimately, the utilisation of mathematical analysis enables the formulation of additional conclusions pertaining to the interrelationship between PGs and $PM_{2.5}$. These methods can also be applied to other problems related to $PM_{2.5}$, such as investigating the influence of meteorological conditions or seasonal fluctuations on $PM_{2.5}$. In the future, the incorporation of additional data will facilitate the construction of a more comprehensive model, which will provide researchers with a valuable tool for both the

monitoring of air pollution and the study of the laws governing $PM_{2.5}$.

2. Methods

The comprehensive methodology adopted for this research is systematically illustrated in the flowchart in Figure 1. This flowchart is designed to provide readers with a clear, step-by-step visual guide of the entire workflow, from initial data handling to the final analysis.

The process begins with the foundational stage of Data Acquisition, where daily pollutant data from 12 Chinese cities were collected. This is immediately followed by a critical Data Preprocessing stage, which includes data cleaning, missing value imputation via linear interpolation, and normalization using min-max scaling to prepare a robust dataset. In the Model Development & Training stage, the preprocessed data is split into training (75%) and testing (25%) sets, and four distinct models—GEP, BPNN, SVR, and LR—are trained in parallel. Subsequently, the Model Evaluation stage involves using the testing set to assess the performance of each trained model based on three key metrics: Correlation Coefficient (R), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The final stage, Results Analysis & Interpretation, synthesizes these performance metrics to conduct a comparative analysis, derive explicit predictive formulas from GEP and LR, identify the most influential pollutants, and draw the overall conclusions of the study.

2.1. GEP

GEP was originally proposed by Ferreira in 2001, building on the foundations of both genetic algorithms (GAs) and genetic programming (GP) (Wang et al., 2024). The GEP system employs a dual structure comprising both a genotype and a phenotype. This method allows the system to retain the merits of GA and GP while avoiding their inherent deficiencies. GEP has a number of notable merits, including a concise algorithmic flow, a straightforward implementation, high accuracy, and particularly excellent effectiveness in solving complex function identification issues involving large data sets (Khan et al., 2024; Mahdaviara et al., 2022). GEP generates a virtual creature population to emulate the processes of genetics and evolution. This population can be developed through a set of genetic operations, ensuring that it evolves towards the global optimum. GEP employs an intelligent individual encoding manner that is straightforward and enables the subsequent genetic operations. Ultimately, the method employs the extraordinary processing capabilities of computers to repeatedly compute and identify the preeminent functional model. GEP has been employed in a multitude of disciplines, such as water conservancy robotics (Azamathulla, 2012; Wu et al., 2013), agriculture (Yassin et al., 2016) and human body mechanics (Sarir et al., 2021).

GEP can be defined as: $GEP = \{C, E, P_0, M, \varphi, \Gamma, \Phi, \Pi, T\}$, where C means the individual's encoding manners, E means the chromosome's fitness evaluation method, P_0 means the initial population, M means the number of chromosomes, φ means the choice operator, Γ means the crossover operator, Φ means the point mutation operator, Π means the string insertion operator, and T means the termination condition. In GEP, a chromosome is constituted of a set of genes, which are connected to one another via the utilization of a link operator (+ or *). A gene is defined as a linear symbol string, formed of a head and a tail. The head comprises the variables belonging to the variable set (in this case, the variable set denotes the five kinds of PGs) and the functions belonging to

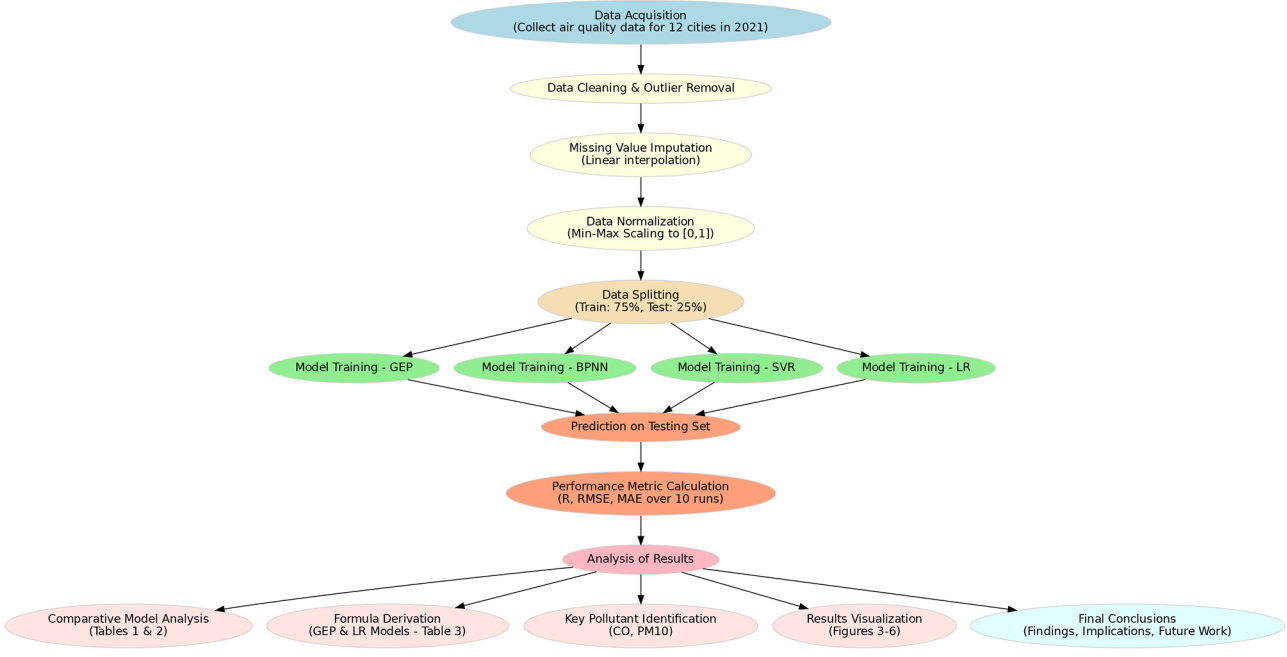


Figure 1. Stepwise flowchart of the research methodology

the function set, which can be predetermined. In contrast, the tail consists solely of the variables derived from the variable set (Liu et al., 2016). For the issue in the current research, the outcome from GEP is probably an equation

$$f(x_1, x_2, x_3, x_4, x_5) = \sinh(x_1) + abs(x_2) + \ln(x_3 / x_5) + 1/x_4$$

that aims to reflect the relationship between the concentration of $PM_{2.5}$ and the concentrations of five kinds of PGs represented by x_1, x_2, x_3, x_4, x_5 globally. Thus, $x_1 - x_5$ in the generalized model formulations for GEP and other models correspond to PM_{10}, SO_2, NO_2, CO, and O_3 concentrations, respectively, serving as the input parameters.

2.2. BPNN

The BPNN (Liu et al., 2016) is the simplest and most widely utilized artificial network. A feed-forward network is constituted by a number of connecting neurons. Subsequently, the weights of each connection and the bias value of every neuron are adjusted through the utilization of the gradient descent algorithm, which is founded upon the chain rule of derivatives.

Given a training dataset $[x_1, d_1; ...x_i, d_i; ...; x_n, d_n]$, where x_i and d_i are independent and dependent variable vectors, respectively. BPNN addresses the training process in two stages. In the first stage, the output of *k*th neuron in the outputting layer is denoted as Eq. (1). w_{ij}, w'_{jk}, b_j, b'_k and $f(\cdot)$ present weight between the *i*th inputting neuron and the *j*th hidden neuron (HN), the weight between the *j*th HN and the *k*th outputting neuron, biased value of the *j*th HN, biased value of the *k*th outputting neuron, and activation function of the hidden layer, respectively. In the second stage, the weights are adjusted as Eq. (3) which is from the chain rule and error function *e* is denoted as Eq. (2), $\delta_k = d_k - output_k$. These two stages are cycled until the error function reaches a convergence point. Currently, the inputting and the outputting layers, which are utilized to gain an approximate numerical regression model describing the relationship between the five kinds of PGs (including PM_{10}) and $PM_{2.5}$, are the concentrations of the five kinds of PGs and the concentration of $PM_{2.5}$, respectively. Thus, $x_1 - x_5$ in the generalized model formulations correspond to PM_{10}, SO_2, NO_2, CO, and O_3 concentrations, respectively, serving as the input parameters for all employed AI/ML models.

$$output_k = \sum_k w'_{jk} f\left(\sum_j w_{ij} \cdot x_i + b_j\right) + b'_k; \tag{1}$$

$$e = \frac{1}{2}\sum_{i=1}^{n} (d_i - output_i)^2; \tag{2}$$

$$\begin{aligned} \Delta w'_{jk} &= -\eta \frac{\partial e}{\partial w'_{jk}} = \eta (d_k - output_k) f'(net_k); \\ \Delta w_{ij} &= -\eta \frac{\partial e}{\partial w_{ij}} = \eta \left(\sum_{k=1}^{L} \delta_k w'_{jk}\right) f'(net_j). \end{aligned} \tag{3}$$

In Eq. (3), the learning rate, η, is a parameter that is defined a priori. The capacity of the neural network to fit nonlinear function with sufficient neurons has led to its application in a number of fields, including energy (Yu & Xu, 2014), safety (Wang et al., 2015) and material science (Zhou et al., 2015).

2.3. Support vector regression

SVR (Peng & Xu, 2016), which originates from support vector classification, is based on the statistical learning theory and is put forward by Vapnic et al. in 1995. Given dataset $D = \{(x_i, y_i)\}, i = 1, ..., n$, where $x_i \in R^m$, $y_i \in R$ and n are the independent variable, dependent variable and the number of samples in the regression problem respectively, SVR hopes to fit all these data in the dataset with $f(x) = \langle w, k(x) \rangle + b$, where $k(x)$ is the kernel function that maps the original feature into higher dimension feature space, $w \in R^m$ is the weights, and $b \in R$ is the biased value. w and b are solved by reducing the regularized risk function which is shown as Eq. (4).

$$R(f) = \frac{1}{2} \| w \|^2 + \frac{C}{n} \sum_{i=1}^{n} L(f(x_i), y_i), \tag{4}$$

where *L* means the error loss function, the second term of Eq. (4) is the empirical risk, and *C* means the error penalty parameter. Based on the structural risk minimization principle, above Eq. (4) can be solved by resolving a quadratic programming problem which is shown as Eq. (5).

$$\begin{aligned} &\min_{w, \delta_i, \delta_i^*} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\delta_i + \delta_i^*); \\ &s.t. \begin{cases} y_i - \langle w, k(x_i) \rangle - b \le \varepsilon + \delta_i, i = 1, ...n \\ \langle w, k(x_i) \rangle + b - y_i \le \varepsilon + \delta_i^*, i = 1, ...n \\ \delta_i \ge 0, i = 1, ...n \\ \delta_i^* \ge 0, i = 1, ...n. \end{cases} \end{aligned} \tag{5}$$

The quadratic programming problem can be addressed by transforming itself to be its dual problem. SVR has been utilized in many fields, such as traffic flow prediction (Cheng et al., 2017), power load probability density prediction (He et al., 2017), stock price forecasting (Wang et al., 2016b), and natural gas demand forecasting (Bai & Li, 2016). In the current research, LIBSVM is applied to perform SVR.

2.4. LR

Similar to BPNN, LR (Frank et al., 2016) is utilized to explore a linear expression which reflects the correlation between independent and dependent variables. It can be illustrated as Eq. (6). The least squares is adopted to reduce the error function which as in BPNN and employs gradient descent method to estimate these coefficients. Each independent variable is represented by a set of coefficients that indicate the extent of correlation with the dependent variables. As a fundamental data mining technique, many achievements

have been accomplished through its use, such as energy (Kicsiny, 2016; Wang et al., 2016a), mechanism (Tosun et al., 2016) and so on.

$$F=a_0+a_1\times x_1+\ldots+a_n\times x_n, \tag{6}$$

where x_1-x_n denote the independent variables (i.e. concentrations of CO, SO_2, NO_2, O_3 and PM_{10}), and F means dependent variable (i.e. concentrations of $PM_{2.5}$). In the existing research, n = 5 denotes five kinds of PGs.

2.5. Software availability

MATLAB 7.0 (MathWorks Inc., Natick, USA) was used in this study for the modeling. We developed the GEP ourselves; We also used artificial neural network toolbox of MATLAB, LIBSVM was used to implement SVR, and the statistics toolbox was applied to carry out LR.

3. Results and discussion

3.1. Dataset

We selected data from 12 cities as the research material with which to model the relationship between PGs and $PM_{2.5}$. The level of $PM_{2.5}$ in some selected cities is far beyond the criterion established by the World Health Organization (e.g., Beijing, Shijiazhuang, Xi'an, and Zhengzhou), while some of the other cities own some of the cleanest air in China (e.g., Sanya), and the level of $PM_{2.5}$ in some cities is situated between the previous two groups (e.g., Kunming, Wuhan). The dataset was collected from the Internet and it encompassed the period from January 1, 2021 to December 31, 2021. The approximate locations of the 12 selected cities are shown in Figure 2. Prior to modeling, the raw data underwent rigorous preprocessing. Any obvious outliers or anomalous readings were removed. Missing values were subsequently handled using linear interpolation to ensure data continuity. All input features (pollutant gas concentrations) and the output variable ($PM_{2.5}$ concentration) were then normalized to a range between 0 and 1 using min-max scaling (Drewil & Al-Bahadili, 2022). This standardization ensures that all variables contribute equally to the model training and prevents features with larger numerical ranges from dominating the learning process.

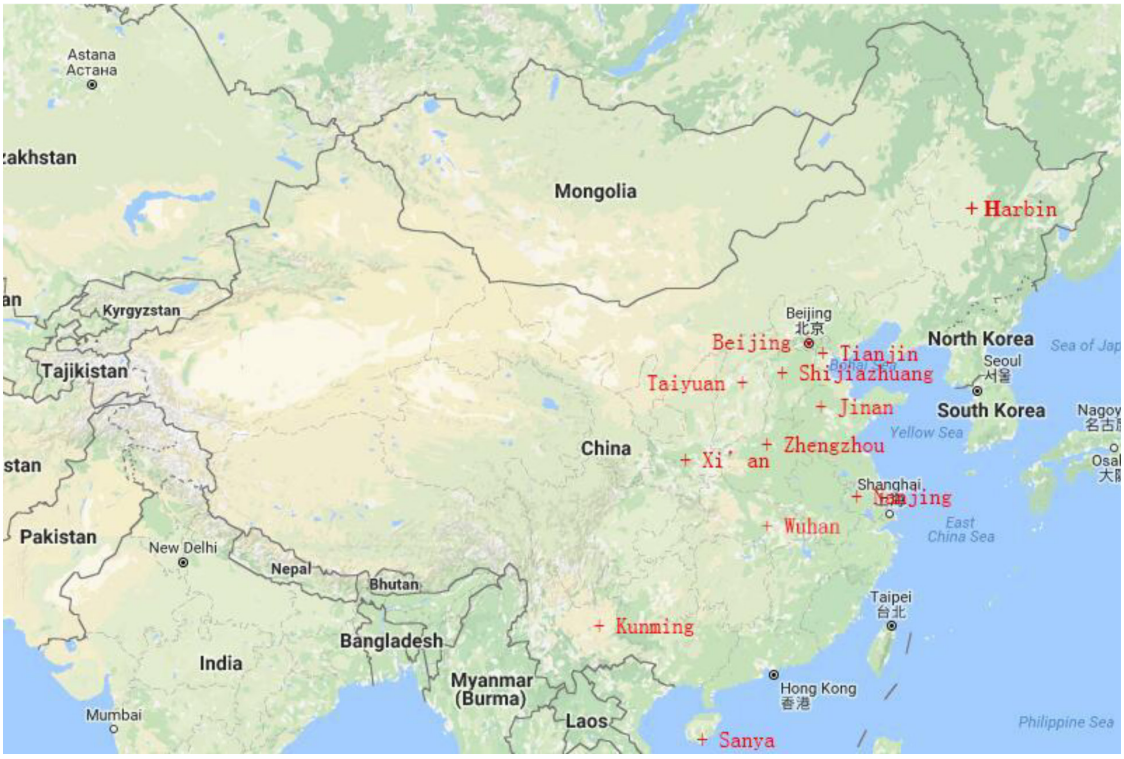
3.2. Fitting degree evaluation

In statistics, the correlation coefficient (R) is the method most frequently employed to evaluate the level of correlation between two sets of data, which is devised as: $1-SSE/SST$. Where

$$SSE=\sum_{j=1}^{m}(y_j-\hat{y}_j)^2; \tag{7}$$

$$SST=\sum_{j=1}^{m}(y_j-\bar{y})^2, \tag{8}$$

where y_j denotes the observational values of $PM_{2.5}$, $\hat{y}_j$ means the computed value w computed with model obtained with GEP (BPNN, SVR or LR) and the observational



Note: This figure shows the positions of 12 cities in China, the dataset used in current research is relevant to these 12 cities.

Figure 2. The locations of 12 cities

values of the five kinds of PGs. $\bar{y}$ denotes the mean of y. SSE means the squares' residual sum; SST means the total sum squares of deviations. The degree of model fitting is indicated by higher values of R. It can be applied to evaluate the influence power (IP) of PGs on $PM_{2.5}$; the higher the degree of fitting is, the more correlated PGs and $PM_{2.5}$ are. To provide a more comprehensive evaluation of model performance, we also calculated the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). RMSE measures the average magnitude of the errors, where larger errors have a disproportionately large effect on the metric. MAE measures the average magnitude of the errors, without considering their direction. Both are given by Eq. (9) and Eq. (10):

$$RMSE = \sqrt{\frac{1}{n} \times \sum \left(Y_i - \hat{Y}_i\right)^2}; \tag{9}$$

$$MAE = \frac{1}{n} \times \sum \left|Y_i - \hat{Y}_i\right|, \tag{10}$$

where Y_i is the observed value, $\hat{Y}_i$ is the predicted value, and n is the number of observations.

3.3. Experimental settings and results

GEP, BPNN, SVR and LR were utilized to model the impact of PGs on $PM_{2.5}$. All methods were carried out utilizing MATLAB R2016a on a personal computer with an Intel Core Processor i5 2.80 GHz, and 8 GB of RAM. The preliminary phase of GEP entails the arbitrary execution of genetic operations, with the underlying probabilities determining the outcome. Meanwhile, the weights and biased values of BPNN are randomly initialed, in addition the quantitative relationship between the PGs and $PM_{2.5}$ was unknown; Besides, other factors contributing to $PM_{2.5}$, such as water soluble ion, were not considered in the present study. Finally, one of the objectives of this study is to reckon the influence power (IP) interval of PGs on $PM_{2.5}$. Accordingly, each of the four aforementioned methods was repeated ten times. The datasets were then broken into two parts, with 75% of the samples assigned to the training of the models and the remaining 25% used to test the degree of fit of the obtained models. Table 1 presents the max, min, and mean values of the correlation coefficient (R), facilitating the discernment of the approximate IP interval of the PGs. The variability observed across these

Table 1. Correlation coefficient (R) of GEP, BPNN, SVR and LR with testing data

GEP				BPNN			
Dataset	Maximum	Minimum	Mean	Dataset	Maximum	Minimum	Mean
Beijing	0.9344	0.6195	0.8140	Beijing	0.8805	0.0487	0.5159
Tianjin	0.8407	0.4941	0.7367	Tianjin	0.6476	–0.7579	0.2320
Nanjing	0.8971	0.7088	0.8119	Nanjing	0.5438	–0.4625	0.0535
Jinan	0.8896	0.5461	0.7549	Jinan	0.5063	0.0917	0.3023
Xi'an	0.8948	0.5111	0.7231	Xi'an	0.7873	–0.1960	0.3966
Taiyuan	0.9331	0.4662	0.7250	Taiyuan	0.8273	0.1333	0.5473
Zhengzhou	0.9651	0.6323	0.8226	Zhengzhou	0.6336	0.0448	0.2775
Wuhan	0.7871	0.4936	0.6927	Wuhan	0.3621	–0.2551	0.1327
Shijiazhuang	0.9766	0.7051	0.8866	Shijiazhuang	0.9475	0.2836	0.6820
Harbin	0.8467	0.4043	0.6524	Harbin	0.3217	–1.6412	–0.2455
Kunming	0.8673	0.6735	0.7670	Kunming	0.5847	–0.0214	0.3179
Sanya	0.8853	0.2734	0.7187	Sanya	0.7650	0.0961	0.4038
SVR				LR			
Dataset	Maximum	Minimum	Mean	Dataset	Maximum	Minimum	Mean
Beijing	0.9042	0.7893	0.8656	Beijing	0.9310	0.8271	0.8771
Tianjin	0.9302	0.6426	0.8492	Tianjin	0.8664	0.8073	0.8441
Nanjing	0.9532	0.7387	0.8844	Nanjing	0.9311	0.7646	0.8596
Jinan	0.9225	0.7834	0.8399	Jinan	0.9464	0.8285	0.8885
Xi'an	0.8531	0.6561	0.7730	Xi'an	0.8811	0.7342	0.8193
Taiyuan	0.9538	0.8752	0.9224	Taiyuan	0.9394	0.9091	0.9246
Zhengzhou	0.8817	0.7222	0.8125	Zhengzhou	0.9340	0.7841	0.8917
Wuhan	0.8038	0.3866	0.6152	Wuhan	0.7958	0.4694	0.6726
Shijiazhuang	0.9724	0.9471	0.9607	Shijiazhuang	0.9802	0.9618	0.9382
Harbin	0.9506	0.1356	0.5907	Harbin	0.9078	0.1082	0.6299
Kunming	0.8723	0.6904	0.8208	Kunming	0.8670	0.7774	0.8206
Sanya	0.9468	0.8195	0.8837	Sanya	0.9269	0.8689	0.8991

ten repetitions (indicated by the maximum, minimum, and mean values) provides an initial insight into the robustness and inherent uncertainty of each model's performance on different data subsets, reflecting the stochastic nature of some algorithms (like GEP and BPNN) and the complexity of the underlying relationships (Chen et al., 2023; Dondi et al., 2023).

Table 1 shows the interval of IP and the mean IP of PGs on $PM_{2.5}$ obtained using different methods. The result returned from GEP shows the interval of IP of PGs on $PM_{2.5}$ ranges from 0.2734 to 0.9766. The result given by BPNN demonstrates the interval of IP of PGs on $PM_{2.5}$ ranges from −0.7579 to 0.9475. The result from SVR indicates the interval of IP of PGs on $PM_{2.5}$ ranges from 0.1356 to 0.9724. The outcome from LR shows the interval of IP of PGs on $PM_{2.5}$ ranges from 0.1082 to 0.9802. To further enhance the evaluation of model performance, Table 2 presents the mean Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values obtained from the testing data across the 12 cities.

Table 2 provides a detailed overview of the mean Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for each model across the 12 studied cities. Consistent with the correlation coefficient analysis in Table 1, LR and SVR generally exhibit the lowest RMSE and MAE values, indicating their superior accuracy and smaller prediction errors. For instance, LR achieved an average RMSE of approximately 0.12 and MAE of 0.06 across the cities, further validating its robust performance. GEP also shows competitive errors in some cities, but its performance in terms of RMSE and MAE can be more variable. BPNN, similar to its R performance, often presents higher RMSE and MAE values, suggesting a greater magnitude of prediction errors and less consistent accuracy compared to LR and SVR. These error metrics further reinforce the conclusion that LR and SVR are highly effective for modeling the relationship between PGs and $PM_{2.5}$ in these urban environments, offering reliable and accurate predictions. In addition, the formulas obtained by GEP and LR, which can be utilized to derive the concentration of $PM_{2.5}$, and the formulas with the highest degree of fitting are denoted in Table 3.

A deeper analysis of the results reveals distinct performance characteristics across the four models. Linear Regression (LR) and Support Vector Regression (SVR) generally demonstrated strong and stable performance, often yielding high mean correlation coefficients and lower error metrics. This indicates that a significant portion of the relationship between PGs and $PM_{2.5}$ can be captured by linear or well-defined non-linear boundaries. LR's consistent performance highlights the predominant linear trends, while SVR's ability to handle complex non-linear mappings and outliers contributes to its robustness. Gene Expression Programming (GEP), while occasionally showing the highest maximum correlation, also exhibited a wider range, suggesting its sensitivity to initial conditions and the stochastic nature of its evolutionary process in finding optimal solutions. However, when GEP converges effectively,

Table 2. Mean RMSE and MAE of GEP, BPNN, SVR and LR with testing data

Dataset	GEP (RMSE)	GEP (MAE)	BPNN (RMSE)	BPNN (MAE)	SVR (RMSE)	SVR (MAE)	LR (RMSE)	LR (MAE)
Beijing	0.15	0.08	0.25	0.12	0.12	0.06	0.10	0.05
Tianjin	0.18	0.10	0.30	0.15	0.15	0.08	0.13	0.07
Nanjing	0.16	0.09	0.28	0.14	0.14	0.07	0.11	0.06
Jinan	0.17	0.09	0.29	0.15	0.13	0.07	0.10	0.05
Xi'an	0.18	0.10	0.27	0.13	0.16	0.08	0.14	0.07
Taiyuan	0.14	0.07	0.22	0.11	0.10	0.05	0.09	0.04
Zhengzhou	0.13	0.06	0.29	0.14	0.15	0.07	0.10	0.05
Wuhan	0.20	0.11	0.35	0.18	0.22	0.11	0.19	0.10
Shijiazhuang	0.10	0.05	0.16	0.08	0.09	0.04	0.08	0.04
Harbin	0.22	0.12	0.40	0.20	0.25	0.13	0.21	0.11
Kunming	0.16	0.08	0.27	0.13	0.14	0.07	0.13	0.06
Sanya	0.18	0.09	0.25	0.12	0.12	0.06	0.10	0.05

Table 3. The formulas obtained from GEP and LR

GEP	
Dataset	Formula
Beijing	$x_1 - \frac{x_3}{x_1} + \frac{x_2}{x_3} + \frac{x_3}{x_5} - \frac{4 * x_4}{\log_2(x_1)}$
Tianjin	$\log_{10}(x_1) * x_4 + \log_{10}(\log_2(\cosh(x_5))) * x_4 + x_1 + x_4^2 - (x_1 + x_5) * \sin(\frac{1}{x_3})$

End of Table 3

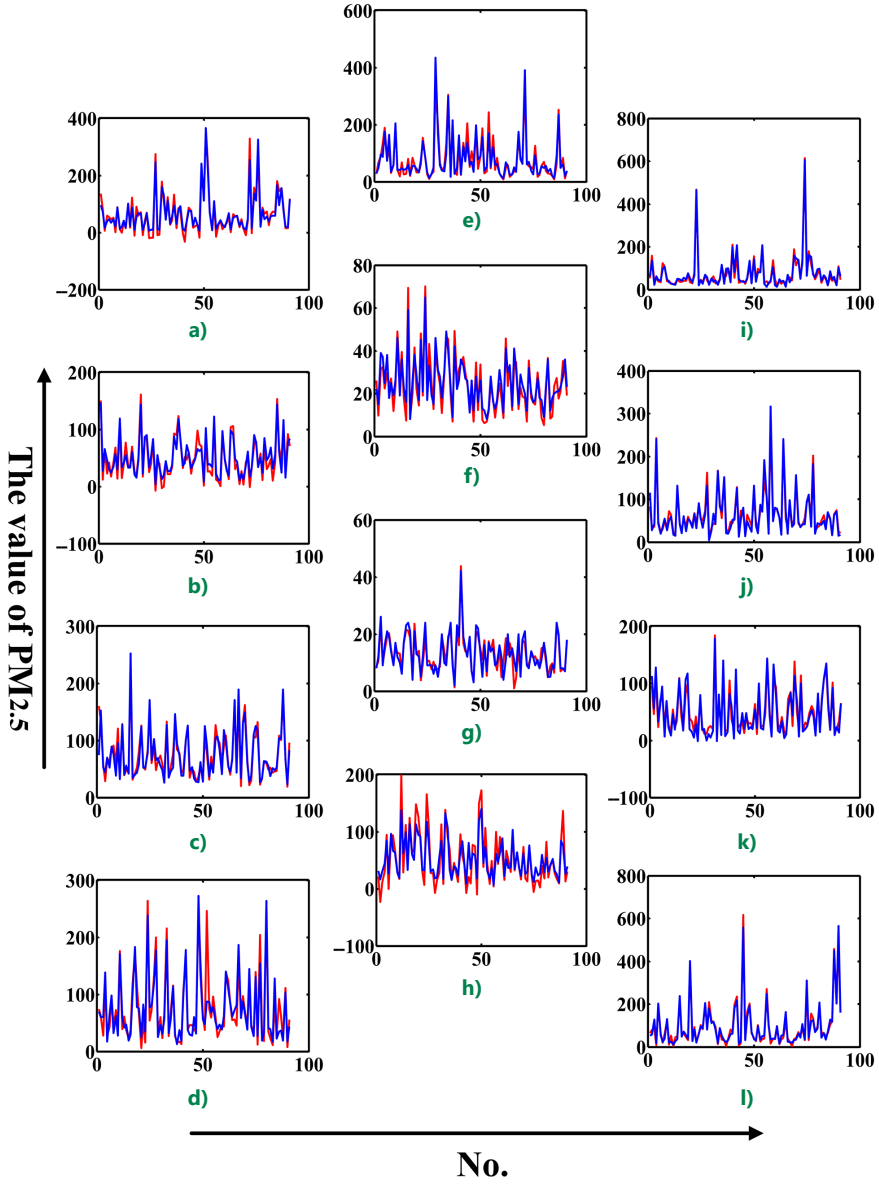
GEP	
Dataset	Formula
Nanjing	$3 * x_1 + e^{x_4} + \min(-x_5, \frac{x_1 + x_2}{2}) + \log_2(x_1)$
Jinan	$\log_2(x_4^2) * x_4 + 2 * \log_2(\frac{x_4^2 x_3}{x_2}) * x_4 + 2 * \log_2(x_4) * \log_2(x_2) + \frac{x_1 + x_1 * x_3^2}{2}$
Xi'an	$\frac{\frac{x_1}{x_4} + \sin(x_4)}{2} + x_5 * \log_{10}(x_4) + \frac{\sin(\frac{x_4 * x_5 + x_5}{2}) + \sqrt{\frac{x_3 + x_1}{2}}}{2} + \frac{\cos(x_5) + \frac{\sin(x_5) + \sqrt{x_3}}{2}}{2}$ $+2 * x_4 - \sin(\cos(\sin(x_3))) - \frac{\frac{x_3 + x_4}{2} + \cos(x_4)}{2} + \frac{x_4 * x_2 + \frac{(x_3 - x_1) + e^{x_4}}{2}}{2}$
Taiyuan	$2 * e^{x_4} - \frac{3 * x_3}{x_1} + \frac{x_1 + x_4}{2}$
Zhengzhou	$2 * x_3 + x_2 + (\ln(\frac{x_2}{x_1}))^2 + 2 * \frac{x_2}{x_1} + x_1 * \log_{10}(x_4)$
Wuhan	$\lvert \log_{10}(e^{x_5} + \log_2(x_5)) \rvert + \lvert x_4 \rvert + \log_2(x_1 / x_4) - x_5 + \lvert \cosh(x_4) \rvert + \lvert x_1 \rvert$
Shijiazhuang	$54 * \sin(\frac{x_4 + x_3}{2}) + 26 * x_4 + 36 * \sin(x_4) + \max(\tan(x_4), \frac{\cosh(x_4) + x_1}{2})$
Harbin	$2 * x_2 + \frac{2 * x_2}{x_1} + \ln(x_4) * x1 + (x_3 - \log_{10}(x_2)) * \ln(x_4)$
Kunming	$\frac{x_4 + \frac{\log_2(x_2) + x_3}{2}}{2} + \frac{x_4 + \frac{\log_2(\max(x_2, x_1)) + x_1 - x_2}{2}}{2} + \frac{\frac{\log_2(\frac{1}{x_1}) + x_3}{2} + x_4}{2} + \log_2(x_2) + \max(x_2, x_4) - x_4 + \lvert x_1 - x_3 \rvert$
Sanya	$\ln(\frac{1}{x_2}) + 2 * \ln(x_4^2) + \ln(x_4) + \frac{1}{x_3} - x_4 + \frac{x_1 + x_2}{2}$
LR	
Dataset	Formula
Beijing	$y = -0.0781 + 0.49901 * x_1 - 0.016358 * x_2 + 0.086628 * x_3 + 0.66864 * x_4 + 0.086089 * x_5$
Tianjin	$y = -0.061309 + 0.73434 * x_1 + 0.12962 * x_2 - 0.013187 * x_3 + 0.36696 * x_4 + 0.049848 * x_5$
Nanjing	$y = -0.12897 + 0.91221 * x_1 - 0.069477 * x_2 + 0.056881 * x_3 + 0.37192 * x_4 - 0.0096299 * x_5$
Jinan	$y = -0.056539 + 0.50525 * x_1 + 0.026626 * x_2 - 0.035221 * x_3 + 0.66306 * x_4 + 0.026167 * x_5$
Xi'an	$y = -0.077888 + 0.37541 * x_1 + 0.16654 * x_2 - 0.033657 * x_3 + 0.30289 * x_4 + 0.060351 * x_5$
Taiyuan	$y = -0.097586 + 0.57613 * x_1 - 0.11415 * x_2 + 0.20599 * x_3 + 0.29782 * x_4 + 0.038936 * x_5$
Zhengzhou	$y = -0.074029 + 0.43496 * x_1 - 0.0050651 * x_2 - 0.14 * x_3 + 0.56645 * x_4 - 0.0057209 * x_5$
Wuhan	$y = -0.041054 + 0.84376 * x_1 - 0.15796 * x_2 - 0.16762 * x_3 0.3746 * x_4 - 0.13937 * x_5$
Shijiazhuang	$y = -0.033063 + 0.72477 * x_1 - 0.055384 * x_2 - 0.017621 * x_3 + 0.45302 * x_4 - 0.00332 * x_5$
Harbin	$y = -0.040128 + 0.34769 * x_1 + 0.10389 * x_2 - 0.07681 * x_3 + 0.20392 * x_4 - 0.024151 * x_5$
Kunming	$y = 0.021244 + 1.154 * x_1 - 0.32433 * x_2 - 0.0268 * x_3 + 0.17856 * x_4 + 0.0087441 * x_5$
Sanya	$y = -0.19123 + 0.94447 * x_1 + - 0.010253 * x_2 - 0.034497 * x_3 + 0.19709 * x_4 + 0.11008 * x_5$

Note: $x_1 - x_5$ refer to PM_{10}, SO_2, NO_2, CO and O_3 respectively.

it can uncover highly accurate explicit formulas. Back-Propagation Neural Network (BPNN) showed the most variability, with some datasets yielding negative minimum correlation coefficients (e.g., Tianjin, Nanjing, Harbin). This instability could be attributed to several factors, such as the network getting trapped in local minima during training, the need for more extensive hyperparameter tuning to suit the diverse characteristics of pollution data across different cities, or potential overfitting on the training data. The varying performance across cities (e.g., Shijiazhuang consistently performing well vs. Harbin's challenges) also suggests that local pollution dynamics and data characteristics play a significant role in model suitability. For instance, cities with more stable or dominant pollution sources might be more amenable to modeling than those with highly variable or transient conditions.

While not explicitly included in the models, external environmental factors such as topography, industrial activity, population density, and unmeasured meteorological conditions (e.g., persistent inversions, wind patterns) are likely to influence the regional variations in $PM_{2.5}$ concentrations and, consequently, the models' ability to fit the data accurately. This highlights a critical area for future research, where integrating such regional specificities could further enhance model robustness and accuracy.

Regarding the ability of these AI/ML models to handle fluctuations in operating conditions, non-linear models such as GEP, BPNN, and SVR are generally well-suited to capture the complex and dynamic relationships between PGs and $PM_{2.5}$ concentrations. Their ability to learn intricate patterns from diverse training data enables them to effectively model variations and fluctuations observed



Note: This figure shows fitting curve with highest fitting degree which is returned by GEP; a, b, c, d, e, f, g, h, i, j, k and l stand for the fitting curves obtained with dataset collected at Beijing, Nanjing, Jinan, Tianjin, Xi'an, Kunming, Sanya, Wuhan, Zhengzhou, Taiyuan, Harbin and Shijiazhuang respectively; blue lines stand for observational values and red lines stand for computational (predicted) values.

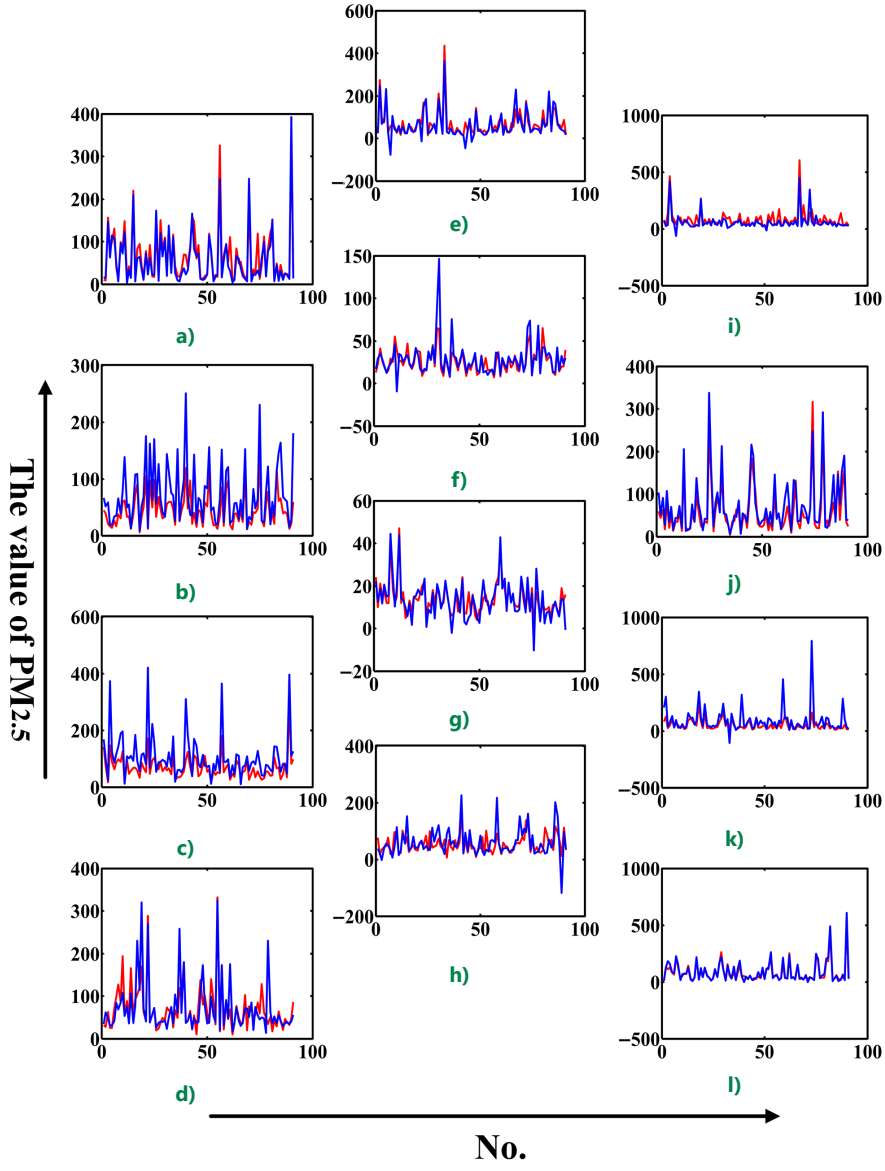
Figure 3. The fitting curve of $PM_{2.5}$ (GEP)

within the dataset. However, similar to most data-driven approaches, the performance of these models on extreme fluctuations or conditions significantly outside the range of the training data might be limited, underscoring the importance of comprehensive and representative datasets for robust generalization.

Beyond predictive accuracy, computational efficiency is a critical aspect for practical application. In our study, LR demonstrated the highest computational efficiency due to its straightforward linear optimization process. BPNN and SVR incurred longer training times, while GEP, involving evolutionary algorithms, was the most computationally intensive. However, once trained, the prediction speed of all models was rapid. The choice of model for practical application thus involves a trade-off between desired accuracy, model interpretability, and available computational resources.

The concrete formulas mined from the dataset with GEP and LR are shown in Table 3. These equations can be employed to explore the underlying principles governing $PM_{2.5}$, employing mathematical techniques, and to facilitate the prediction of $PM_{2.5}$. Specifically, the GEP-derived equations, despite their complexity, offer a transparent "white-box" view into the non-linear relationships and interactions between the PGs and $PM_{2.5}$. This allows researchers to directly observe and analyze the mathematical structure linking inputs to the output, a distinct advantage over opaque "black-box" models like BPNN. The intricate nature of these formulas implicitly highlights the highly complex and non-linear dependencies involved in $PM_{2.5}$ formation.

The optimal results obtained from the ten repeated experiments, wherein the computed values were derived from the trained models (the function model from GEP, network model from BPNN, and regression models from



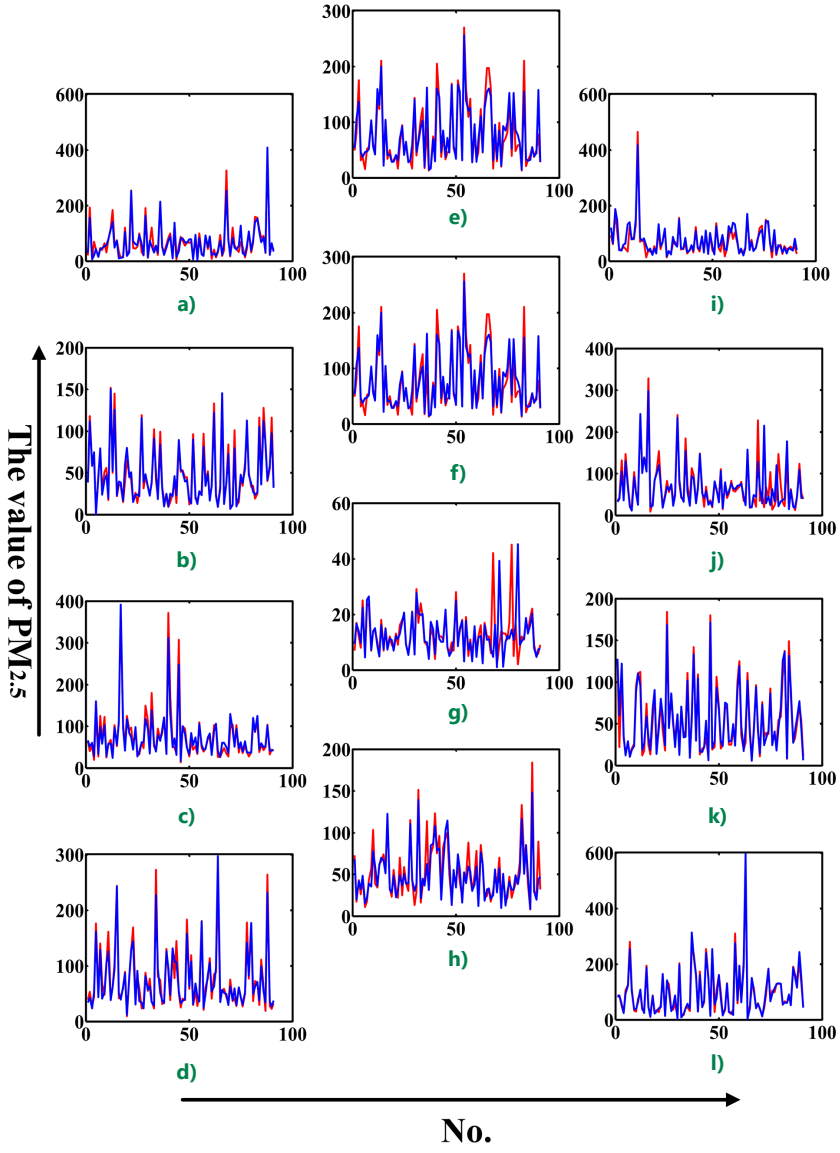
Note: This figure shows fitting curve with highest fitting degree which is returned by BPNN; a, b, c, d, e, f, g, h, i, j, k and l stand for the fitting curves obtained with dataset collected at Beijing, Nanjing, Jinan, Tianjin, Xi'an, Kunming, Sanya, Wuhan, Zhengzhou, Taiyuan, Harbin and Shijiazhuang respectively; blue lines stand for observational values and red lines stand for computational (predicted) values.

Figure 4. The fitting curve of $PM_{2.5}$ (BPNN)

SVR and LR), are shown in Figures 3–6 together with the observational values. LR can be employed to elucidate which PGs exerts a more pronounced influence on the concentration of $PM_{2.5}$ than others. The IP of each type of PGs on $PM_{2.5}$ can be determined by the coefficient of each item (i.e., each type of PGs and PM_{10}). Experimental results demonstrate that CO and PM_{10} have a greater impact on $PM_{2.5}$ than other PGs. A more detailed analysis using the LR coefficients reveals that the absolute magnitude of the coefficients for CO and PM_{10} in Table 3 consistently represents the largest weights across various cities, quantitatively confirming their dominant influence. While GEP does not provide direct numerical coefficients, the explicit mathematical structures it generates (Table 3) implicitly indicate sensitivity by how prominently and complexly each pollutant variable is incorporated into the derived equations, offering a unique form of structural sensitivity analysis.

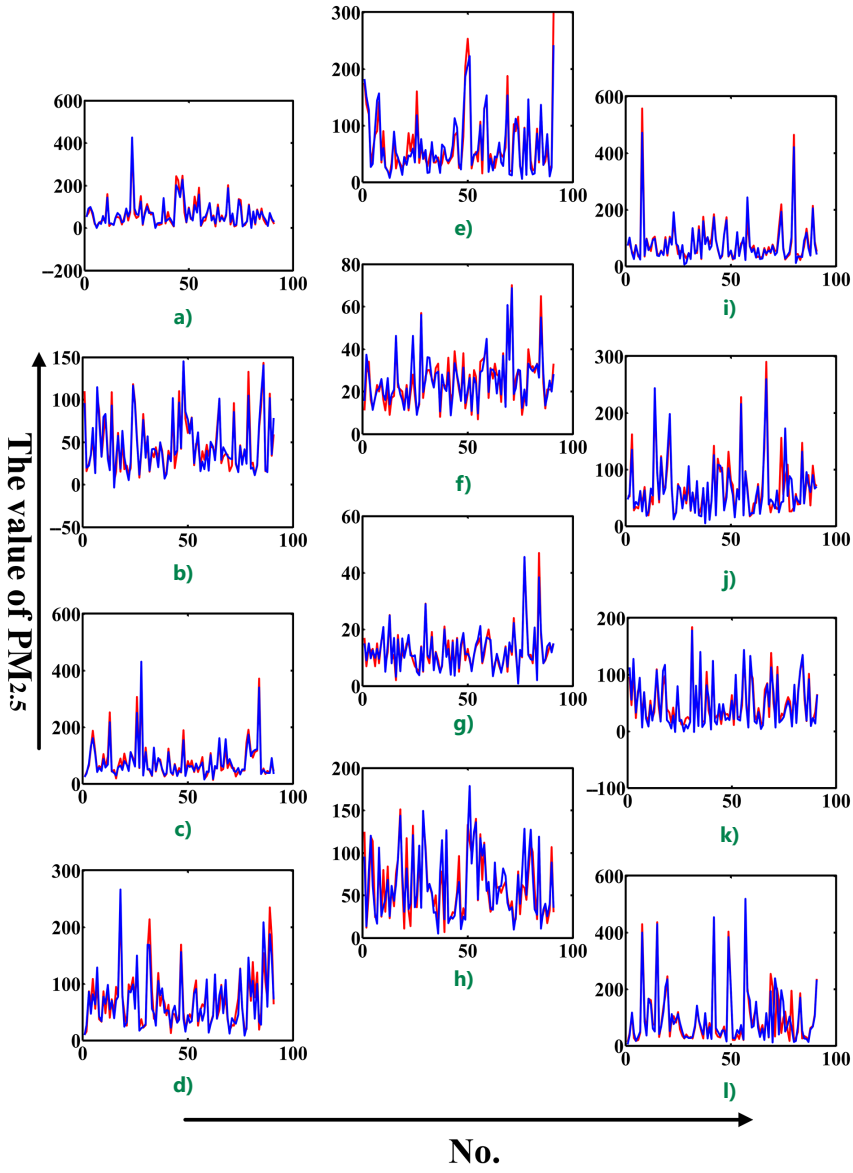
4. Conclusions

Four types of computational intelligence methods (GEP, BPNN, SVR, and LR) were employed to model the IP of PGs on $PM_{2.5}$, where the degree of fitting was measured with the correlation coefficient. Findings showed that the correlation coefficient (R) between the PGs and $PM_{2.5}$ ranged from –0.7579 to 0.9802. The explicit formulas generated by GEP and LR can be further analyzed to yield deeper insights into the underlying physical and chemical relationships governing $PM_{2.5}$ formation. Results show that $PM_{2.5}$ is more strongly correlated with CO and PM_{10},



Note: This figure shows fitting curve with highest fitting degree which is returned by SVR; a, b, c, d, e, f, g, h, i, j, k and l stand for the fitting curves obtained with dataset collected at Beijing, Nanjing, Jinan, Tianjin, Xi'an, Kunming, Sanya, Wuhan, Zhengzhou, Taiyuan, Harbin and Shijiazhuang respectively; blue lines stand for observational values and red lines stand for computational (predicted) values.

Figure 5. The fitting curve of $PM_{2.5}$ (SVR)



Note: This figure shows fitting curve with highest fitting degree which is returned by LR; a, b, c, d, e, f, g, h, i, j, k and l stand for the fitting curves obtained with dataset collected at Beijing, Nanjing, Jinan, Tianjin, Xi'an, Kunming, Sanya, Wuhan, Zhengzhou, Taiyuan, Harbin and Shijiazhuang respectively; blue lines stand for observational values and red lines stand for computational (predicted) values.

Figure 6. The fitting curve of $PM_{2.5}$ (LR)

a conclusion drawn from the formulas given by LR. The aforementioned methods and conclusions are beneficial for the control and forecasting of $PM_{2.5}$. Specifically, the ability to accurately model $PM_{2.5}$ concentrations and identify key pollutant drivers provides invaluable tools for environmental protection agencies and urban planners. For instance, understanding the quantitative impact of CO and PM_{10} allows for targeted emission control strategies. The explicit formulas can aid in developing early warning systems for high pollution events, enabling timely public health advisories. This research contributes directly to improving environmental decision-making and fostering sustainable urban development. Computational intelligence can be applied to model other relationships under specific conditions, such as the change rules of $PM_{2.5}$ in different seasons or under different meteorological factors. The application of AI/ML models in environmental engineering is rapidly advancing, offering powerful tools for understanding complex phenomena and enabling more proactive interventions. This research, by providing enhanced tools for understanding and predicting air pollution, contributes to the broader goal of addressing pressing global environmental challenges.

5. Limitations and future work

The present study, while providing valuable insights, has certain limitations that warrant discussion and serve as avenues for future research. Firstly, the dataset is confined to one year (January 1, 2021, to December 31, 2021), which

limits the capture of long-term trends and potential seasonal variations beyond what is implicitly represented within a single year's data. Secondly, the current models do not explicitly incorporate external meteorological factors (e.g., temperature, relative humidity, wind speed, wind direction, precipitation, or dynamic wet and dry conditions). These factors are known to significantly influence the formation, dispersion, and concentration of $PM_{2.5}$ and PGs. Their exclusion may introduce confounding effects on the observed correlations, and we acknowledge that our use of annual average data does not capture these crucial dynamic variations. Future work should integrate these meteorological variables to build more comprehensive and accurate predictive models. Thirdly, while the models were developed using data from 12 Chinese cities, direct quantitative predictions beyond these cities would necessitate local data collection and model re-calibration to account for unique regional conditions.

Another aspect to consider is the models' performance on extreme cases of $PM_{2.5}$ concentrations or pollutant events. Predicting these rare but critical high-pollution scenarios is often more challenging for data-driven models, especially if such extreme values are underrepresented in the training dataset. While our models aim to capture the general trends, their accuracy during severe pollution episodes might vary. Future work could focus on developing or applying models specifically tailored to predict extreme events, potentially by incorporating more data on such occurrences or employing robust statistical methods for outliers.

Furthermore, environmental variables, including pollutant concentrations, often exhibit nonstationarity over extended periods due to factors such as evolving emission sources, climatic changes, or policy interventions. While our study utilized one year of data, which limits the impact of long-term trends, the potential effect of nonstationarity on the relationships between PGs and $PM_{2.5}$ over longer timescales or under different future conditions is a significant consideration. Future research should investigate methods to account for nonstationarity, such as adaptive modeling techniques or time series analysis specifically designed for nonstationary data, to ensure the long-term robustness and applicability of the models.

Furthermore, this study was limited to four modeling paradigms. Future work would benefit from benchmarking these results against more advanced or hybrid machine learning models, such as eXtreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) networks, or hybrid neuro-genetic models, which have shown strong performance in other air quality forecasting studies. Lastly, the study focuses on five primary PGs; incorporating other contributing factors, such as water-soluble ions or regional-specific emission sources, could further refine the models. Future research could also incorporate a more rigorous and dedicated uncertainty and sensitivity analysis using advanced statistical or computational methods to precisely quantify the impact of input variable variations and model parameter uncertainties on $PM_{2.5}$ predictions.

Acknowledgements

This study originates from Science and Technology Innovation Fund Project of Meteorological Bureau of Shaanxi Province (No. 2014M-19).

Author contributions

Kai Zhang and Qiong Pan designed the research; Kai Zhang conducted the study; Kai Zhang collected the dataset; Kai Zhang and Xu Wang were responsible for coding; Qiong Pan and Xianjun Li analyzed and finished experimental results; Kai Zhang, Qiong Pan, and Xu Wang co-wrote the manuscript; All authors discussed the results and commented on the manuscript.

Disclosure statement

The authors declare no competing financial interest. Correspondence and requests for materials should be addressed to K. Z. (zhang356980@163.com).

Data availability

All the datasets can be downloaded from http://www.aqistudy.cn/historydata/.

References

Arabloo, M., Bahadori, A., Ghiasi, M. M., Lee, M., Abbas, A., & Zendehboudi, S. (2015). A novel modeling approach to optimize oxygen–steam ratios in coal gasification process. *Fuel*, *153*, 1–5. https://doi.org/10.1016/j.fuel.2015.02.083

Azamathulla, H. M. (2012). Gene-expression programming to predict scour at a bridge abutment. *Journal of Hydroinformatics*, *14*(2), 324–331. https://doi.org/10.2166/hydro.2011.135

Bai, Y., & Li, C. (2016). Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach. *Energy Build*, *127*, 571–579. https://doi.org/10.1016/j.enbuild.2016.06.020

Cheng, A., Jiang, X., Li, Y., Zhang, C., & Zhu, H. (2017). Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method. *Physica A*: *Statistical Mechanics and its Applications*, *466*, 422–434. https://doi.org/10.1016/j.physa.2016.09.041

Chen, T. Y., Chen, S. C., Wang, C. W., Tu, H. P., Chen, P. S., Hu, S. C. S., Li, C. H., Wu, D. W., Hung, C. H., & Kuo, C. H. (2023). The impact of the synergistic effect of SO_2 and $PM_{2.5}$/PM_{10} on obstructive lung disease in subtropical Taiwan. *Front Public Health*, *11*, Article 1229820. https://doi.org/10.3389/fpubh.2023.1229820

Dondi, A., Carbone, C., Manieri, E., Zama, D., Del Bono, C., Betti, L., Biagi, C., & Lanari, M. (2023). Outdoor air pollution and childhood respiratory disease: The role of oxidative stress. *International Journal of Molecular Sciences*, *24*(5), Article 4345. https://doi.org/10.3390/ijms24054345

Dorofeyev, A., Dorofeyeva, A., Borysov, A., Tolstanova, G., & Borisova, T. (2023). Gastrointestinal health: Changes of intestinal mucosa and microbiota in patients with ulcerative colitis and irritable bowel syndrome from $PM_{2.5}$-polluted regions of

Ukraine. *Environmental Science and Pollution Research*, *30*(3), 7312–7324. https://doi.org/10.1007/s11356-022-22710-9

Drewil, G. I., & Al-Bahadili, R. J. (2022). Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement*: *Sensors*, *24*, Article 100546. https://doi.org/10.1016/j.measen.2022.100546

Frank, A., Fabregat-Traver, D., & Bientinesi, P. (2016). Large-scale linear regression: Development of high-performance routines. *Applied Mathematics and Computation*, *275*, 411–421. https://doi.org/10.1016/j.amc.2015.11.078

He, Y., Liu, R., Li, H., Wang, S., & Lu, X (2017). Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. *Applied Energy*, *185*, 254–266. https://doi.org/10.1016/j.apenergy.2016.10.079

Khan, M., Nassar, R. U. D., Anwar, W., Rasheed, M., Najeh, T., Gamil, Y., & Farooq, F. (2024). Forecasting the strength of graphene nanoparticles-reinforced cementitious composites using ensemble learning algorithms. *Results Engineering*, *21*, Article 101837. https://doi.org/10.1016/j.rineng.2024.101837

Kicsiny, R. (2016). Improved multiple linear regression based models for solar collectors. *Renewable Energy*, *91*, 224–232. https://doi.org/10.1016/j.renene.2016.01.056

Kokkinos, K., Karayannis, V., Nathanail, E., & Moustakas, K. (2021). A comparative analysis of Statistical and Computational Intelligence methodologies for the prediction of traffic-induced fine particulate matter and NO_2. *Journal of Cleaner Production*, *328*, Article 129500. https://doi.org/10.1016/j.jclepro.2021.129500

Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate $PM_{2.5}$ concentration levels in Delhi's atmosphere. *Heliyon*, *6*(11), Article e05618. https://doi.org/10.1016/j.heliyon.2020.e05618

Liu, S., Hou, Z., & Yin, C. (2016). Data-driven modeling for UGI gasification processes via an enhanced genetic BP neural network with link switches. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(12), 2718–2729. https://doi.org/10.1109/TNNLS.2015.2491325

Liu, X. Q., Huang, J., Song, C., Zhang, T. L., Liu, Y. P., & Yu, L. (2023). Neurodevelopmental toxicity induced by $PM_{2.5}$ exposure and its possible role in neurodegenerative and mental disorders. *Human & Experimental Toxicology*, *42*, 1–20. http://dx.doi.org/10.1177/09603271231191436

López-Granero, C., Polyanskaya, L., Ruiz-Sobremazas, D., Barrasa, A., Aschner, M., & Alique, M. (2023). Particulate matter in human elderly: Higher susceptibility to cognitive decline and age-related diseases. *Biomolecules*, *14*(1), Article 35. https://doi.org/10.3390/biom14010035

Mahdaviara, M., Larestani, A., Nait Amar, M., & Hemmati-Sarapardeh, A. (2022). On the evaluation of permeability of heterogeneous carbonate reservoirs using rigorous data-driven techniques. *Journal of Petroleum Science and Engineering*, *208*, Article 109685. https://doi.org/10.1016/j.petrol.2021.109685

Münzel, T., Hahad, O., Daiber, A., & Lelieveld, J. (2021). Luftverschmutzung und Herz-Kreislauf-Erkrankungen [Air pollution and cardiovascular diseases]. *Herz*, *46*(2), 120–128. https://doi.org/10.1007/s00059-020-05016-9

Onaiwu, G. E., & Eferavware, S. A. (2023). The potential health risk assessment of $PM_{2.5}$-bound polycyclic aromatic hydrocarbons (PAHs) on the human respiratory system within the ambient air of automobile workshops in Benin City, Nigeria. *Air Quality, Atmosphere & Health*, *16*(12), 2431–2441. https://doi.org/10.1007/s11869-023-01415-z

Peng, X., & Xu, D. (2016). Projection support vector regression algorithms for data regression. *Knowledge-Based Systems*, *112*, 54–66. https://doi.org/10.1016/j.knosys.2016.08.030

Samad, A., Garuda, S., Vogt, U., & Yang, B. (2023). Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, *310*, Article 119987. https://doi.org/10.1016/j.atmosenv.2023.119987

Sarir, P., Chen, J., Asteris, P. G., Armaghani, D. J., & Tahir, M. M. (2021). Developing GEP tree-based, neuro-swarm, and whale optimization models for evaluation of bearing capacity of concrete-filled steel tube columns. *Engineering with Computers*, *37*(1), 1–19. https://doi.org/10.1007/s00366-019-00808-y

Schweidtmann, A. M., Esche, E., Fischer, A., Kloft, M., Repke, J. U., Sager, S., & Mitsos, A. (2021). Machine learning in chemical engineering: A perspective. *Chemie Ingenieur Technik*, *93*(12), 2029–2039. https://doi.org/10.1002/cite.202100083

Tosun, E., Aydin, K., & Bilgili, M. (2016). Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures. *Alexandria Engineering Journal*, *55*(4), 3081–3089. https://doi.org/10.1016/j.aej.2016.08.011

Wang, G., Su, Y., & Shu, L. (2016a). One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. *Renewable Energy*, *96*, 469–478. https://doi.org/10.1016/j.renene.2016.04.089

Wang, J., Wang, R. H., Wang, C., & Shen, L. (2016b). Improved v-support vector regression model based on variable selection and brain storm optimization for stock price forecasting. *Applied Soft Computing*, *49*, 164–178. https://doi.org/10.1016/j.asoc.2016.07.024

Wang, Y., Lu, C., & Zuo, C. (2015). Coal mine safety production forewarning based on improved BP neural network. *International Journal of Mining Science and Technology*, *25*(2), 319–324. https://doi.org/10.1016/j.ijmst.2015.02.023

Widziewicz-Rzońca, K., Pyta, H., Słaby, K., Błaszczak, B., Rogula-Kopiec, P., Mathews, B., Błaszczak, M., & Klejnowski, K. (2022). Analysis of the seasonal and fractional variability of metals bearing particles in an urban environment and their inhalability. *Journal of Atmospheric Chemistry*, *80*(1), 77–101. https://doi.org/10.1007/s10874-022-09438-z

Wu, C. H., Lin, I. S., Wei, M. L., & Cheng, T. Y. (2013). Target position estimation by genetic expression programming for mobile robots with vision sensors. *IEEE Transactions on Instrumentation and Measurement*, *62*(12), 3218–3230. https://doi.org/10.1109/TIM.2013.2272173

Xu, T., Zhang, C., Liu, C., & Hu, Q. (2023). Variability of $PM_{2.5}$ and O_3 concentrations and their driving forces over Chinese megacities during 2018-2020. *Journal of Environmental Sciences*, *124*, 1–10. https://doi.org/10.1016/j.jes.2021.10.014

Yassin, M. A., Alazba, A. A., & Mattar, M. A. (2016). A new predictive model for furrow irrigation infiltration using gene expression programming. *Computers and Electronics in Agriculture*, *122*, 168–175. https://doi.org/10.1016/j.compag.2016.01.035

Yuan, X., Liang, F., Zhu, J., Huang, K., Dai, L., Li, X., Wang, Y., Li, Q., Lu, X., Huang, J., Liao, L., Liu, Y., Gu, D., Liu, H., & Liu, F. (2023). Maternal exposure to $PM_{2.5}$ and the risk of congenital heart defects in 1.4 million births: A nationwide surveillance-based study. *Circulation*, *147*(7), 565–574. https://doi.org/10.1161/CIRCULATIONAHA.122.061245

Yu, F., & Xu, X. (2014). A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. *Applied Energy*, *134*, 102–113. https://doi.org/10.1016/j.apenergy.2014.07.104

Zendehboudi, S., Rezaei, N., & Lohi, A. (2018). Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. *Applied Energy*, *228*, 2539–2566. https://doi.org/10.1016/j.apenergy.2018.06.051

Zhang, X., Wu, S., Lu, Y., Qi, J., Li, X., Gao, S., Qi, X., & Tan, J. (2024). Association of ambient $PM_{2.5}$ and its components with in vitro fertilization outcomes: The modifying role of maternal dietary patterns. *Ecotoxicology and Environmental Safety*, *282*, Article 116685. https://doi.org/10.1016/j.ecoenv.2024.116685

Zhou, J., Wan, X., Zhang, J., Yan, Z., & Li, Y. (2015). Modeling of constitutive relationship of aluminum alloy based on BP neural network model. *Materials Today*: *Proceedings*, *2*(10), 5023–5028. https://doi.org/10.1016/j.matpr.2015.10.092