Journal of Electronic Research and Application, 2025, Volume 9, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on an Air Pollutant Data Correction Method Based on Bayesian Optimization Support Vector Machine

Xingfu Ou[1], Miao Zhang[2], Wenfeng Chen[1]*

[1]Institute of Intelligent Manufacturing, Foshan Polytechnic, Foshan 528137, Guangdong, China
[2]College of Automation, Guangdong University of Technology, Guangzhou 510006, Guangdong, China

**Author to whom correspondence should be addressed.*

Abstract: Miniature air quality sensors are widely used in urban grid-based monitoring due to their flexibility in deployment and low cost. However, the raw data collected by these devices often suffer from low accuracy caused by environmental interference and sensor drift, highlighting the need for effective calibration methods to improve data reliability. This study proposes a data correction method based on Bayesian Optimization Support Vector Regression (BO-SVR), which combines the nonlinear modeling capability of Support Vector Regression (SVR) with the efficient global hyperparameter search of Bayesian Optimization. By introducing cross-validation loss as the optimization objective and using Gaussian process modeling with an Expected Improvement acquisition strategy, the approach automatically determines optimal hyperparameters for accurate pollutant concentration prediction. Experiments on real-world micro-sensor datasets demonstrate that BO-SVR outperforms traditional SVR, grid search SVR, and random forest (RF) models across multiple pollutants, including $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3. The proposed method achieves lower prediction residuals, higher fitting accuracy, and better generalization, offering an efficient and practical solution for enhancing the quality of micro-sensor air monitoring data.

Keywords: Air quality monitoring; Data calibration; Support vector regression; Bayesian optimization; Machine learning

Online publication: August 7, 2025

1. Introduction

With the accelerated development of global urbanization and industrialization, air pollution has evolved into a severe global environmental challenge, seriously restricting the sustainable development of cities and posing an increasingly serious threat to human health. In order to achieve high-density, real-time, and refined monitoring of urban air quality, in recent years, micro air quality detectors have attracted extensive attention and application from researchers and environmental protection agencies around the world due to their significantly low cost, flexible deployment characteristics, and potential for high spatial resolution monitoring [1]. This type of innovative equipment can not only efficiently monitor fine particulate matter ($PM_{2.5}$), inhalable particulate matter (PM_{10}), and six typical air pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and ozone (O_3) in the environment, but also simultaneously collect multimodal meteorological data such as temperature, humidity, wind speed, air pressure, precipitation, etc. These rich data dimensions provide unprecedented support for the construction of an urban air quality perception network with a wide coverage and finer granularity, making up for the limitation of insufficient spatial coverage of traditional fixed monitoring stations.

Despite the significant advantages of micro detectors, they still face many technical and application challenges in actual operation. Compared with the professional-grade instruments used in national or regional high-precision reference monitoring stations, micro-detectors have significant gaps in basic sensor performance, long-term operational stability, cross-sensitivity, and adaptability to complex environments. The original monitoring data are generally affected by multiple complex factors such as the inherent accuracy limitations of sensors, drastic changes in ambient temperature and humidity, cross-reactions between different pollutants, and sensor aging and drift, resulting in widespread systematic deviations and high uncertainty in the data. This data quality problem makes it difficult for the original output of micro air quality detectors to directly meet the rigor of scientific research, the decision-making support needs of policy making, and the accuracy requirements of public services. Therefore, how to effectively and robustly correct the monitoring data of micro devices and keep them highly consistent with the reference-level monitoring results has become a key bottleneck problem to give full play to the application value of micro air quality detectors and incorporate them into the mainstream air quality management system.

In view of this, in recent years, academia and industry have invested a lot of energy in trying to introduce a variety of data-driven methods to correct the output of micro detectors. The commonly used correction methods in current research mainly include linear models based on statistical regression (such as linear regression and polynomial regression) and nonlinear models based on machine learning (such as random forest (RF), gradient boosting tree (GBDT), and deep neural network (DNN)) [2]. The above methods have improved the bias and uncertainty of micro air quality detector data to a certain extent, but they still have their own limitations. For example, linear models are difficult to capture complex nonlinear relationships; some nonlinear models such as random forests may face the challenge of dimensionality disaster when processing high-dimensional data; and although neural network models have strong fitting ability, they usually require a large amount of data for training, and the model has poor interpretability. At the same time, they have high requirements for computing resources and may not be suitable for edge deployment scenarios of micro sensors. In addition, the performance of these models is highly dependent on the reasonable selection of hyperparameters. Traditional parameter adjustment methods, such as manual experience adjustment or grid search (GS), are inefficient and time-consuming in high-dimensional parameter space, and are very easy to fall into local optimal solutions. It is difficult to fully explore the potential of the model, and it is difficult to adapt to the complex and changeable data characteristics and environmental conditions in practical applications [3,4]. In order to overcome the above challenges and further improve the performance and practicality of the air quality data correction model, this paper innovatively proposes a correction method for support vector machine (BO-SVM) based on Bayesian optimization [5]. The core idea of this method is to use the powerful global optimization ability of the Bayesian optimization (BO) framework. Bayesian optimization globally models the complex hyperparameter space of SVM by constructing surrogate models such as Gaussian process (GP), and adaptively selects the next evaluation point in combination with the acquisition function, so that the optimal hyperparameter combination of the SVM model can be found efficiently and intelligently under a limited number of evaluations. This mechanism can significantly improve the correction accuracy and generalization ability of the SVR (support vector regression, as a specific application of SVM in regression tasks) model, making it more robust when processing micro-monitoring data. This paper will conduct a comprehensive and rigorous experimental verification of the proposed BO-SVM method by pairing it with actual micro-air quality detection data (including $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3 and related meteorological parameters) and high-precision reference station data. The experimental results will show from multi-dimensional evaluation indicators (such as RMSE, MAE, R^2) that the BO-SVM model is superior to the traditional SVR and other common data-driven regression models in multiple key indicators. It has strong practical value and broad promotion potential, and provides a new technical path for improving the quality of micro air quality sensor data.

2. Methods

This study proposes a Bayesian Optimization-Support Vector Regression (BO-SVR) model to perform high-precision correction on the raw data collected by micro air quality monitors. This method cleverly combines the inherent advantages of support vector machines (SVM) in modeling small samples, high dimensions, and nonlinear data, and the ability of Bayesian optimization (BO) to perform efficient global search in complex hyperparameter spaces [6]. Through this integrated strategy, we hope to build a robust and accurate correction model to cope with common challenges in environmental monitoring data, such as noise, multi-source interference, sensor drift, and calibration bias, thereby significantly improving the reliability and accuracy of micro monitoring data.

2.1. Problem definition

The correction of air quality monitoring data is essentially a regression problem. We define the multidimensional feature vector collected by the micro monitor at a specific time t as $x_t \in R_d$. This feature vector comprehensively covers a variety of environmental parameters, including but not limited to:

Particle concentration: fine particulate matter ($PM_{2.5}$), inhalable particulate matter (PM_{10}).

Gaseous pollutant concentrations: carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), etc.

Meteorological variables: ambient temperature, relative humidity, atmospheric pressure, wind speed, wind direction, precipitation intensity, etc. These meteorological factors have a significant impact on the response characteristics of the sensor and the diffusion of pollutants, and are key auxiliary information for building an accurate calibration model.

At the same time, it is assumed that there is a high-precision reference monitoring station that provides a true and reliable air quality measurement value $y_t \in \mathrm{R}$ at the same time t. This reference value usually comes from a national or regional standard monitoring station, and its data quality has been strictly calibrated and quality controlled.

The calibration goal of this study is to learn and construct a nonlinear mapping function $f{:}\mathrm{R}^d \rightarrow \mathrm{R}$, so that the function can accurately map the original reading x_t of the micro-monitor to the true value y_t of the corresponding reference monitoring station. Mathematically, this goal can be expressed as:

$$y_t = f(\mathrm{x}_t) + \varepsilon_t$$

Where ε_t represents the residual term, which includes noise, random errors, and unmodeled system deviations that are not captured by the model. Our core task is to find an optimal function f, so that the norm of the residual term ε_t is as small as possible, so that the constructed regression model $f(x_t)$ can be as close to the reference data y as possible, and finally achieve high-precision and automatic correction of the monitoring value of the micro-device. This process not only improves the individual accuracy of the micro-sensor but also provides data quality assurance for the large-scale, low-cost deployment of the air quality monitoring network.

2.2. Support Vector Regression Model (SVR)

Support Vector Regression (SVR) is an extended application of Support Vector Machine (SVM) in regression tasks. It adheres to the core idea of SVM - Structural Risk Minimization (SRM). Different from the traditional empirical risk minimization (ERM), SRM aims to minimize the training error while ensuring the complexity of the model, thereby improving the generalization ability of the model and effectively avoiding overfitting. This is particularly important for environmental monitoring data with limited samples, high-dimensional features, and nonlinear relationships.

The goal of SVR is to learn a nonlinear regression function that can best fit the training data. Specifically, the function maps the original input data x to a higher-dimensional feature space φ(x), thereby finding an optimal linear regression hyperplane in the high-dimensional space [7]. The mathematical expression of this regression function is:

$$f(\mathrm{x})=\langle \mathrm{w},\phi(\mathrm{x})\rangle +b$$

Where $\phi(\mathrm{x})$ is a nonlinear mapping function that maps the data point x in the original input space to the high-dimensional feature space. w is the weight vector in the high-dimensional space, which defines the normal direction of the regression hyperplane; *b* is the bias term, which determines the position of the hyperplane in the feature space.

The uniqueness of SVR lies in the introduction of the -insensitive loss function. During the training process, errors falling within the ε range (that is, the absolute value of the difference between the predicted value and the true value is less than or equal to ε) are not penalized. This gives SVR robustness to noise, because it allows the model to “tolerate” data fluctuations within a certain error range, focusing on capturing the main trends of the data rather than being sensitive to every tiny error, which is crucial for the task of correcting air quality data containing measurement noise [8–10].

In order to find the optimal regression function *f(x)*, SVR is trained by minimizing the following objective function:

$$\min_{\mathrm{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$

This objective function consists of two parts:

(1) $\frac{1}{2}\|\mathrm{w}\|^2$: Regularization term, used to minimize the norm of the weight vector “w,” thereby maximizing the interval of the regression hyperplane, thereby controlling the complexity of the model, and enhancing the generalization ability of the model.

(2) $C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$: Empirical risk term, where ξ_i and $\xi*_i$ are slack variables, representing the excess error of data point i on the upper and lower sides of the ξ-pipeline, respectively. $C>0$ is a penalty factor used to balance the trade-off between model complexity and training error.

When the *C* value is large, the model will more strictly penalize the error beyond the ε-pipeline, making the model more inclined to fit the training data, but may increase the risk of overfitting.

When the *C* value is small, the model has a higher tolerance for errors, which may lead to a simpler model, but the training error may increase, and there is a risk of underfitting.

The above objective function is subject to the following constraints:

$$\begin{cases} y_i-\langle \ \mathrm{w},\phi(x_i)\rangle \ -b\leq\varepsilon+\xi_i \\ \langle \mathrm{w},\phi(x_i)\rangle+b-y_i\leq\varepsilon+\xi_i^* \\ \qquad \xi_i,\xi_i^* \ \geq 0 \end{cases}$$

The constraints ensure that most data points can fall within the "insensitive area" formed by *f(x)*+*ε*, or are allowed to exceed this area by introducing slack variables ξ_i and ξ^*_i, but the excess amount is penalized.

In order to avoid the huge amount of computation caused by directly calculating ϕ(x) in high-dimensional feature space, SVR uses duality theory and kernel tricks. By introducing Lagrange multipliers, SVR can eventually be transformed into a decision function in the form of the following kernel function:

$$f(\mathrm{x})=\sum_{i=1}^{n}(\alpha_i-\alpha_i^* \ )K(\mathrm{x}_i,\mathrm{x})+b$$

Among them, α_i and α^*_i are Lagrange multipliers. Only those data points located on or outside the boundary of the *ε*-insensitive region (i.e., support vectors) have non-zero α_i–α^*_i values. The support vectors determine the final regression hyperplane. $K\left(\mathrm{x}_i,\mathrm{x}_j\right)=\langle \ \phi(x_i),\phi(x_j)\rangle$ is a kernel function that calculates the inner product in the feature space without explicitly calculating ϕ(x).

In practical applications, the choice of kernel function is crucial to the performance of SVR. This paper selects Gaussian Radial Basis Function (RBF) as the kernel function, and its expression is:

$$K(\mathrm{x}_i,\mathrm{x})=\exp\boxed{f()}(-\gamma\left\|x_i-x_j\right\|^2)$$

Wherein, γ>0 is a key hyperparameter of the RBF kernel function, known as the kernel width parameter, which controls the range of influence of a single training sample on the regression function, thereby determining the nonlinear modeling ability of the model:

The larger the γ value, the faster the decay of the kernel function, the smaller the influence range of the support vector, and the model will be more inclined to capture local data features, resulting in increased model complexity and possible overfitting.

The smaller the γ value, the slower the decay of the kernel function, the larger the influence range of the support vector, and the model will be more inclined to capture global data trends, reducing model complexity and possibly leading to underfitting.

In summary, the performance of the SVR model is highly dependent on the reasonable setting of its key hyperparameters C and γ. Traditional manual parameter adjustment or grid search methods are inefficient and difficult to guarantee the optimal solution, especially in multidimensional hyperparameter space. Therefore, it is necessary to introduce Bayesian optimization to search these hyperparameters efficiently and intelligently.

2.3. Bayesian optimization

Traditional support vector regression (SVR) models usually rely on grid search or manual adjustment to determine the hyperparameters $\theta=(C,\gamma)$, but in high-dimensional hyperparameter space, such methods are not only computationally expensive, but also prone to local optimality. In order to improve the efficiency of parameter search and model performance, this paper introduces the Bayesian Optimization (BO) strategy to perform global modeling and adaptive exploration of the hyperparameter space, so as to automatically find the optimal parameter combination.

Bayesian optimization guides the selection of the next round of evaluation points by constructing a probabilistic proxy model of the objective function $\mathscr{L}(\theta)$. The objective function $\mathscr{L}(\theta)$ is usually defined as the model loss under cross-validation, such as the root mean square error (RMSE) or the mean absolute error (MAE). Its core process includes the following three steps [11]:

(1) Proxy model construction:

Bayesian optimization uses Gaussian process (GP) as the prior proxy model of the objective function, that is:

$$\mathscr{L}(\theta) \sim GP(\mu(\theta),k(\theta,\theta^{'}))$$

where $\mu(\theta)$ represents the mean function and $k(\theta,\theta')$ is the covariance function, which is used to characterize the similarity between different parameters.

(2) Acquisition function guides search:

In order to balance exploration and exploitation, this paper uses expected improvement (EI) as the acquisition function:

$$\mathrm{EI}(\theta)=\mathrm{E}[\max\boxed{f()}(0,f_{best}-f(\theta))]$$

where represents the objective function value of the currently known best point. The EI function can guide the next step of evaluation at the most potential point.

(3) Iterative update and termination condition:

In each round of iteration, select the parameter combination θ that maximizes EI, evaluate it on the true objective function, and update the GP model accordingly. Repeat the above process until the maximum number of iterations is reached or the convergence condition is met.

2.4. BO-SVR air quality correction framework

The BO-SVR air quality data correction model proposed in this paper has an overall process including feature construction, model initialization, objective function setting, and model training and prediction [12], as follows:

(1) Feature construction:

The input feature vector consists of the concentrations of "two particles and four gases" pollutants and five types of typical meteorological variables, including:

$$\mathrm{x}=[\mathrm{PM}_{2.5},\mathrm{PM}_{10},\mathrm{CO},\mathrm{NO}_2,\mathrm{SO}_2,\mathrm{O}_3,T,H,P,WS,R]$$

Where T represents temperature, H represents relative humidity, P represents atmospheric pressure, WS represents wind speed, and R represents precipitation. This multimodal information jointly characterizes the impact of environmental changes on sensor output.

(2) Model initialization:

The initial hyperparameter search space of the support vector regression (SVR) model is set as follows:

$$C\in[10^{-2},10^{3}],\gamma\in[10^{-4},10]$$

Where C is the penalty factor, which controls the balance between fitting and generalization; γ is the parameter of the kernel function, which affects the model's ability to respond to nonlinear features.

(3) Objective function setting:

The objective function of Bayesian optimization is defined as the average root mean square error (RMSE) calculated based on K-fold cross-validation, and the expression is as follows:

$$\mathscr{L}(\theta)=\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(y_i^{(k)}-\hat{y}_i^{(k)})^2}$$

Wherein, $y_i^{(k)}$ and $\hat{y}_i^{(k)}$ represent the true value and predicted value of the i-th sample in the k-th fold, respectively, and n_k is the number of samples in the fold. This objective function serves as an evaluation indicator for Bayesian optimization and guides the search process for the optimal parameters.

(4) Model training and prediction:

Use Bayesian optimization to search for the optimal hyperparameter θ^{*} =(C^{*} ,γ^{*}) in the predefined parameter space, and train the final SVR model based on it. After training, the model can be used to correct the raw pollutant data collected by miniature air quality detection devices, thereby improving its accuracy and consistency, making it closer to the actual measurement results of reference stations or high-precision instruments.

3. Experimental settings and result analysis

3.1. Experimental data and preprocessing

This study selected the synchronous monitoring data of the air quality automatic monitoring station of a certain municipal environmental protection bureau (hereinafter referred to as the “reference station”) and the micro air quality detection equipment deployed within 1 km of its periphery, with a time span from October 2024 to March 2025. Both types of equipment recorded the concentration of “two particles and four gases” pollutants and five types of meteorological variables such as temperature, humidity, wind speed, air pressure, and precipitation, and the data sampling frequency was 1 hour.

In order to improve the model training effect and data consistency, the following preprocessing operations were performed on the original data before the experiment:

Time alignment and missing value interpolation (using linear interpolation);

Outliers (Z-score > 3) were eliminated;

All features were Z-score standardized.

The final experimental data set contains about 4,000 valid samples, divided into a training set and a test set in a ratio of 7:3.

3.2. Comparison method settings

To verify the correction performance of the proposed BO-SVR model, the following four typical models are selected for comparison experiments:

Raw: uncorrected raw micro-device data;

SVR (default parameters): traditional support vector regression model without parameter adjustment;

GS-SVR: SVR model with grid search parameter adjustment;

RF: correction method based on random forest regression;

BO-SVR (method in this paper): SVR model with Bayesian optimization parameter adjustment.

All the above models use the same feature input, and the training set is consistent with the test set to ensure fairness.

3.3. Evaluation indicators

In order to comprehensively measure the prediction accuracy of the model after correction, this paper selects the following three types of evaluation indicators:

Root mean square error (RMSE):

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

Mean absolute error (MAE):

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

Coefficient of determination (R^2):

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Wherein, y_i is the reference station data, $\hat{y}_i$ is the model prediction value, $\bar{y}$ is the reference data mean, and n is the total number of samples.

3.4. Experimental results and analysis

Table 1 lists in detail the performance indicators of the original data (Raw) and four different models (default SVR, grid search optimized SVR (GS-SVR), random forest (RF) and Bayesian optimized SVR (BO-SVR)) on the test set in the $PM_{2.5}$ concentration correction task [13–15], including root mean square error (RMSE), mean absolute error (MAE) and determination coefficient (R^2).

Table 1. Performance of the classic model in $PM_{2.5}$ concentration correction task

Model	RMSE (μg/m^3)	MAE (μg/m^3)	R^2
Raw	9.72	4.63	0.61
SVR (默认)	4.88	3.12	0.74
GS-SVR	3.47	2.06	0.78
RF	2.92	1.61	0.81
BO-SVR	1.36	1.22	0.85

The results show that the original data (Raw) has the highest RMSE and MAE, the lowest R^2, and a significant error. The default SVR, GS-SVR, and RF models are better than the original data in all indicators, which shows the effectiveness of machine learning correction.

The BO-SVR model showed the best performance, with an RMSE of only 1.36 μg/m^3 and a MAE of 1.22 μg/m^3, both of which are the lowest among all models, indicating the smallest prediction error. At the same time, R^2 is as high as 0.85, far exceeding other models, indicating that the degree of fit between its predicted value and the true value is the highest. This fully verifies that Bayesian optimization can significantly improve the correction accuracy and stability of SVR, enabling it to achieve excellent results in $PM_{2.5}$ data correction.

Figure 1 shows the performance of the BO-SVR model in the $PM_{2.5}$ concentration prediction task and the comparison trend with the actual monitoring values of the reference station. As can be seen from the figure, the predicted value (red dotted line) is highly consistent with the true value (blue solid line) in terms of overall trend, and can better capture the temporal variation characteristics of pollutant concentrations, especially in areas with large fluctuations (such as sample indexes 100 to 300, 600 to 800), it still maintains good fitting accuracy without obvious lag or distortion.

In addition, the BO-SVR model has a relatively accurate fitting effect in multiple high-value peak segments, indicating that it has a strong modeling ability when processing nonlinear and highly volatile characteristic data, further verifying the effectiveness of Bayesian optimization in hyperparameter selection, and significantly improving the generalization ability of the model by automatically searching for the optimal C=1.4979 and γ=0.9997.

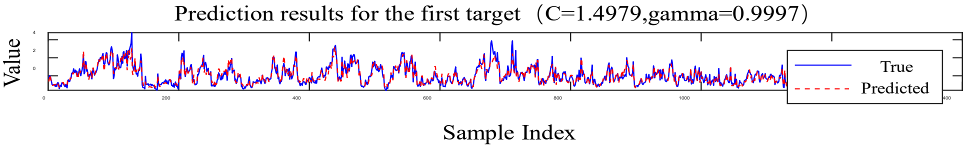


Figure 1. Performance of BO-SVR model in $PM_{2.5}$ concentration prediction task

Figure 2 shows a two-dimensional thermodynamic diagram of the BO-SVR model performance as the penalty factor C and kernel function parameter γ change. The diagram intuitively reveals the interactive effects of hyperparameters on model performance and the optimal region.

As can be seen from the figure, the model performance is poor when both the C value and the γ value are small (the dark blue area in the lower left corner), indicating that there may be underfitting. As the C value and the γ value gradually increase, the model performance is significantly improved, and the color gradually transitions from blue to bright yellow. Specifically, when the C value reaches about 30 or more and the γ value increases to about 4 to 6, the model performance reaches a peak (bright yellow area, performance index as high as 0.94 to 0.96). This shows that the SVR model can best balance the model complexity and fitting ability in this area, and effectively capture the nonlinear characteristics in $PM_{2.5}$ data.

This thermodynamic diagram intuitively verifies the excellent efficiency of Bayesian optimization in finding high-performance hyperparameter combinations. It can intelligently focus on promising parameter areas and avoid blind searches, thereby significantly improving the accuracy and optimization efficiency of the SVR model in $PM_{2.5}$ correction.

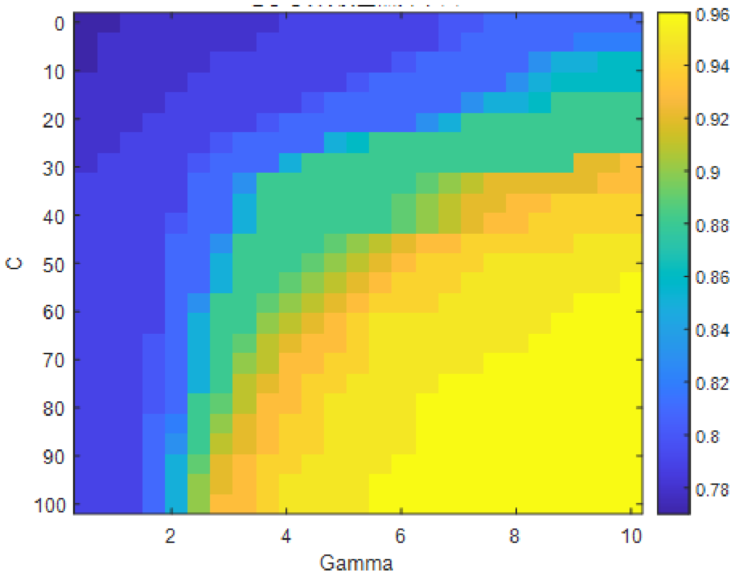


Figure 2. Analysis of BO-SVR model hyperparameter optimization results

Figure 3 shows the distribution histogram of prediction residuals of different correction models in $PM_{2.5}$ data correction (Raw data, SVR, GS-SVR, RF, and BO-SVR). The more concentrated the prediction residuals are at zero and the narrower the distribution is, the higher the correction accuracy of the model is.

As can be seen from the figure, the residual distribution of the original data (Raw, orange) is the widest and the peak is low, indicating that its error is the largest and the volatility is strong. After correction by traditional SVR (blue) and grid search optimized GS-SVR (yellow), the residual distribution converges, the peak increases, and the error decreases. The random forest (RF, purple) model shows a narrower residual distribution and a higher peak, showing its better correction ability.

However, the residual distribution of BO-SVR (green) is the most concentrated, with the highest peak and significantly close to zero. This shows that the SVR model after Bayesian optimization has the smallest prediction error and the highest correction accuracy, effectively reducing the data deviation of the micro-monitor, and verifying the excellent performance of BO-SVR in $PM_{2.5}$ data correction.

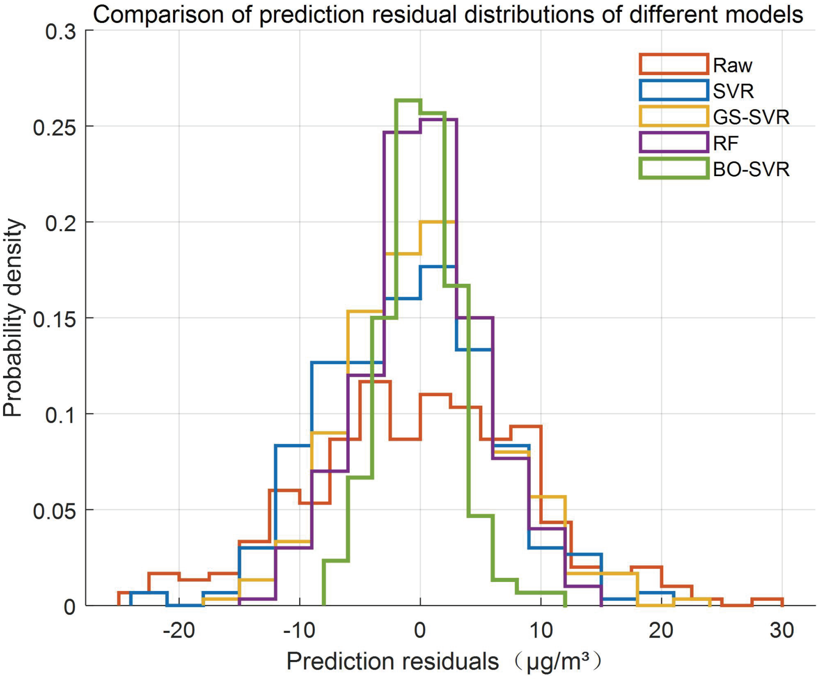


Figure 3. Comparison of prediction residual distribution of different models

Figure 4 shows the box plots of the prediction residuals of $PM_{2.5}$ data for the original data (Raw) and four correction models (SVR, GS-SVR, RF, BO-SVR). The narrower the box, the closer the median is to zero, and the fewer outliers, indicating that the correction effect of the model is better and the error is smaller.

As can be seen from the figure, the box of the original data (Raw) is the widest, the median deviates from zero, and there are many outliers, indicating that its error range is wide and the deviation is large. The box width of the SVR and GS-SVR models has narrowed, and the median is closer to zero, showing the basic correction effect. The box of the random forest (RF) is further narrowed, but there are still some outliers. The box of the BO-SVR model is the narrowest, its interquartile range is the smallest, the median is closest to zero, and the number of outliers is significantly reduced. The results show that the SVR model after Bayesian optimization shows the smallest prediction residual and the highest stability in $PM_{2.5}$ data correction, which significantly improves the accuracy of the data.

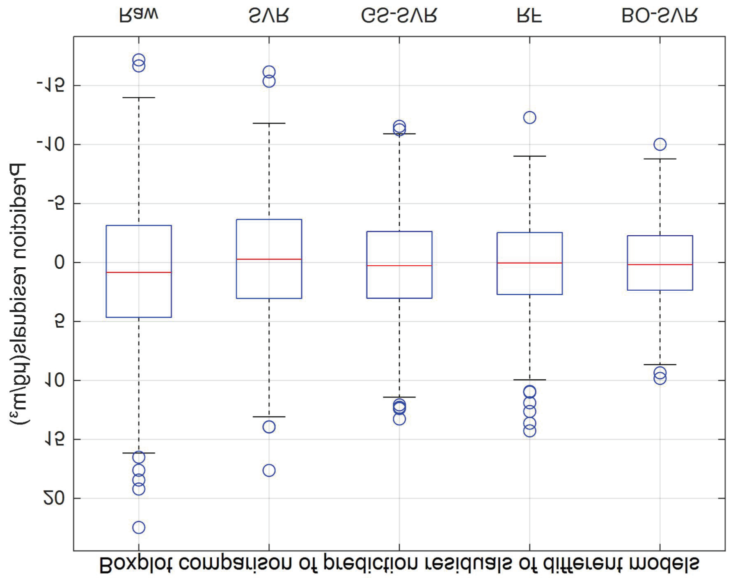
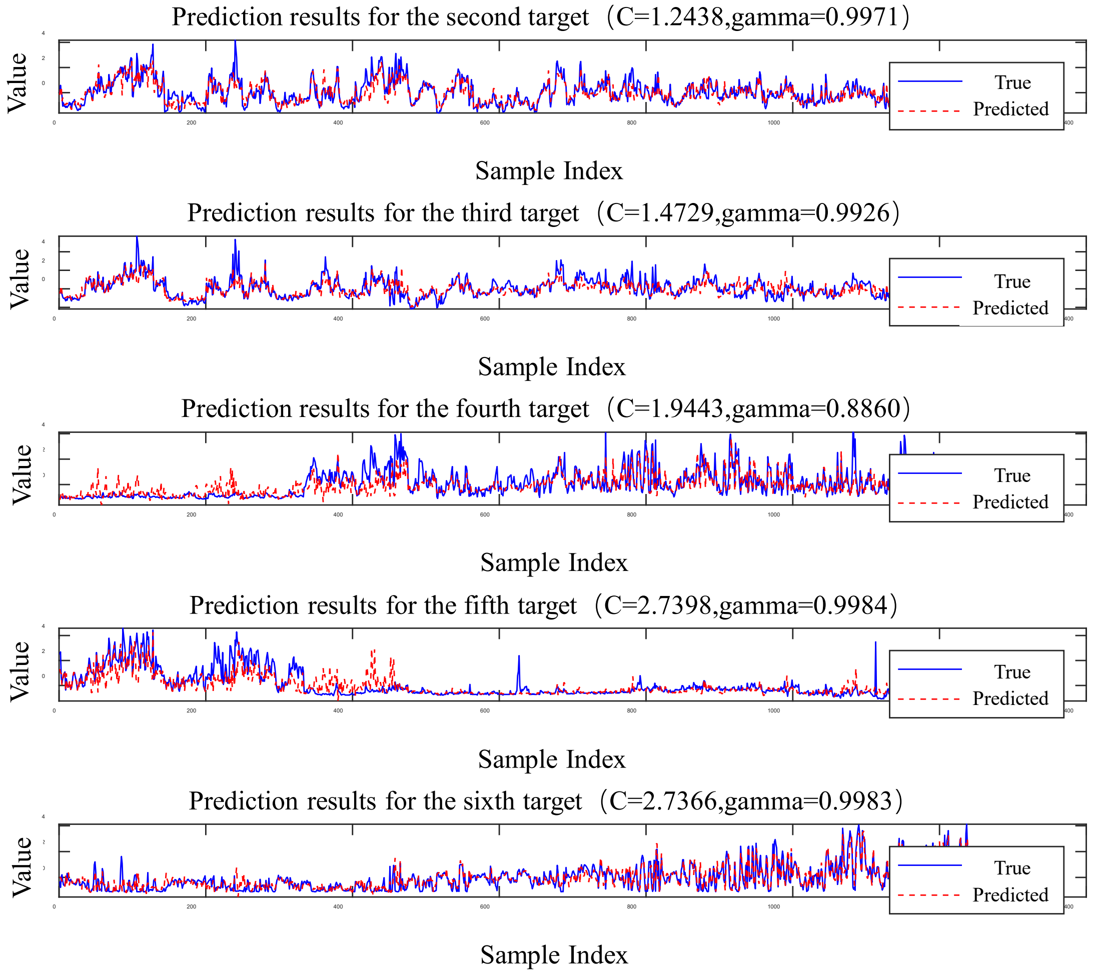


Figure 4. Comparison of error box plots of different models on PM2.5 data

Figure 5 shows the prediction results of the BO-SVR model on the test sets of five major air pollutants: PM_{10}, CO, NO_2, SO_2, and O_3. Each sub-figure compares the true value (blue solid line) with the predicted value of the BO-SVR model (red dotted line), and indicates the corresponding optimized C and γ parameters.

In general, the BO-SVR model shows good correction ability for data of different pollutants. Among the five types of pollutants, PM_{10}, CO, NO_2, SO_2, and O_3, the predicted curve (red dotted line) is highly consistent with the true curve (blue solid line), and the trend tracking ability is strong, indicating that the model can effectively capture the dynamic characteristics of pollutant concentrations over time. Especially at peaks or valleys with large concentration fluctuations, the BO-SVR model can still maintain good prediction accuracy and effectively correct the readings of the micro-monitor to a level close to the true value.

Although the concentration ranges and fluctuation characteristics of different pollutants are different, Bayesian optimization can find the optimal hyperparameter combination for the SVR model to adapt to the respective data characteristics (such as PM_{10} (C=1.2438, γ=0.9971) and O_3 (C=2.7366, gamma=0.9983)), which further confirms the strong advantages of Bayesian optimization in improving the generalization ability and adaptability of the model, enabling it to optimize the SVR model in a targeted manner, thereby showing stable high performance in the multi-pollutant monitoring data correction task.


Prediction results for the second target（C=1.2438,gamma=0.9971）

Prediction results for the third target（C=1.4729,gamma=0.9926）

Prediction results for the fourth target（C=1.9443,gamma=0.8860）

Prediction results for the fifth target（C=2.7398,gamma=0.9984）

Prediction results for the sixth target（C=2.7366,gamma=0.9983）

Figure 5. BO-SVR model correction results for different air pollutants

In summary, the above experimental results consistently show that introducing Bayesian optimization into the support vector regression model can significantly improve the correction accuracy, stability, and model generalization ability of micro air quality monitoring data. BO-SVR not only comprehensively surpasses the existing mainstream correction methods in key performance indicators, but also provides a solid technical guarantee for the accurate application of micro sensor data through an efficient hyperparameter search mechanism.

4. Conclusion and prospects

This study proposes and verifies a Bayesian optimization-based support vector regression (BO-SVR) data correction method to address the accuracy and stability issues of micro air quality monitoring data. By deeply analyzing the hyperparameter optimization process of the BO-SVR model in the $PM_{2.5}$ thermodynamic diagram, we confirm that Bayesian optimization can efficiently and intelligently search for the optimal SVR hyperparameter combination, significantly improving the model performance.

The experimental results show that compared with traditional methods such as raw data, default SVR, grid search SVR, and random forest, the BO-SVR model has excellent advantages in key evaluation indicators such as

RMSE, MAE, and R^2. Its prediction residual distribution is the most concentrated and close to zero, and the error box plot shows the smallest error range and the least outliers, which fully demonstrates the powerful ability of BO-SVR in improving data accuracy. In addition, BO-SVR shows good universality and robustness in the correction of multiple types of pollutants such as PM_{10}, CO, NO_2, SO_2, and O_3, and can effectively cope with the complex correction needs of different pollutants.

In summary, this study provides an efficient and high-precision solution for the correction of micro air quality sensor data, which strongly supports the construction of a high-density, real-time air quality monitoring network. Future work can further explore the fusion correction of multi-source heterogeneous data, online incremental learning methods, and lightweight strategies for deploying models to edge devices to adapt to more complex practical application scenarios.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Castell N, Dauge FR, Schneider P, et al., 2017, Can Commercial Low-Cost Sensor Platforms Contribute to Air Quality Monitoring and Exposure Estimates? Environment International, 99: 293–302.

[2] Zimmerman N, Presto AA, Kumar SPN, et al., 2018, A Machine Learning Calibration Model Using Random Forests to Improve Sensor Performance for Lower-Cost Air Quality Monitoring. Atmospheric Measurement Techniques, 11(1): 291–313.

[3] Jiao W, Hagler G, Williams R, et al., 2016, Community Air Sensor Network (CAIRSENSE) Project: Evaluation of Low-Cost Sensor Performance in a Suburban Environment. Atmospheric Measurement Techniques, 9: 5281–5292.

[4] Liu W, Zhang Z, Zhou S, 2021, A Review of Low-Cost Sensor Data Calibration for Air Quality Monitoring. Journal of Instrumentation, 42(11): 161–169.

[5] Shahriari B, Swersky K, Wang Z, et al., 2016, Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE, 104(1): 148–175.

[6] Snoek J, Larochelle H, Adams RP, 2012, Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems, 25: 2951–2959.

[7] Smola AJ, Scholkopf B, 2004, A Tutorial on Support Vector Regression. Statistics and Computing, 14(3): 199–222.

[8] Duan KB, Keerthi SS, 2005, Which Is the Best Multiclass SVM Method? An Empirical Study. Pattern Recognition, 38(12): 2097–2109.

[9] Zhang Y, Ding Y, Hao H, et al., 2019, Air Quality Forecasting Using a SVR Model Based on Hybrid Optimization Algorithm. Atmospheric Pollution Research, 10(5): 1529–1539.

[10] Liu Z, Liu Y, Yu M, et al., 2022, Improved Support Vector Regression with Multiple Kernel Learning for PM2.5 Prediction. Science of The Total Environment, 802: 149801.

[11] Li H, Zhou T, Zhang J, 2022, Research on Air Quality Prediction Based on Improved Support Vector Regression Model. Environmental Science and Technology, 45(5): 132–139.

[12] Zhang X, Han Z, Cui J, et al., 2020, Air Quality Prediction Model Based on Wavelet Analysis and Support Vector Machine. Systems Engineering and Electronic Technology, 42(2): 410–417.

[13] Wei P, Yang Z, Wang Y, et al., 2021, A Hybrid Calibration Approach for Low-Cost PM2.5 Sensors Using Multiple

Machine Learning Models. Environmental Research, 200: 111355.

[14] Guo H, Zhang X, Wang Y, et al., 2020, Calibration of Low-Cost Air Quality Sensors Using Transfer Learning and Ensemble Learning. Sensors, 20(12): 3502.

[15] Wang H, Hu Q, Liu C, 2021, PM2.5 Concentration Prediction Based on LSTM-SVR Combination Model. Environmental Science and Management, 46(1): 112–116.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.