PM$_{2.5}$ Inversion Based on XGBoost And LightGBM Integrated Models

Yanyou Ren, Yan Zhang*, Shurui Fan
School of Electronic Information Engineering, Hebei University of Technology, Tianjin 300401, China

Abstract: Accurate inversion of PM2.5 concentration is crucial for haze management. Currently commonly used inversion methods cannot accurately invert the concentration in non-site areas, so this paper proposes a PM2.5 concentration inversion method based on an integrated learning model. The method utilises the Top Atmospheric Reflectance (TOAR), observation angle and meteorological element data as input features, and screens the important features by Random Forest, and constructs an integrated inversion model using XGBoost and LightGBM. The results show that the model built by TOAR improves $R^2$ by 2.9% and reduces RMSE and MAE by 2.67 and 1.45, respectively, compared with the AOD-based model, and our model has an inversion accuracy of 0.95, which is better than other models. We used the model to estimate and analyse the historical PM2.5 concentration changes at Huaihe station in Tianjin, China, and the results were consistent with the trend of the actual PM2.5 concentration distribution, and it is clear that the proposed model has a high inversion accuracy.

1. Introduction

Air pollution has been a public health problem on an international scale, the most dangerous being particulate matter with an aerodynamic diameter of less than 2.5 millimetres, known as PM2.5\(^1\).

Traditional atmospheric monitoring programmes have high construction costs and uneven distribution, while satellite remote sensing monitoring data with high spatial resolution and fast speed can make up for the shortcomings of monitoring stations\(^2\). In recent years, researchers and scholars\(^3\) have used different methods to study the link between satellite AOD and ground-based PM2.5 concentrations, mainly including pattern scale factor method, semi-empirical method, statistical modelling method and machine learning method. However, AOD inversion can be limited by the inversion algorithm with great uncertainty\(^4\), and the pixel points of the AOD products are severely missing, which leads to the lack of AOD-based products and the spatial coverage cannot be guaranteed. Statistics show that MODIS AOD can be missing up to 94% in mainland China\(^5\). Later, researcher\(^6\) used the top-of-atmosphere reflectance TOAR to invert PM2.5 concentration, researcher Shen\(^7\) used a deep confidence network, and Wang\(^8\) used an intelligent long- and short-term memory network to construct a link between PM2.5 and TOAR, and both compared with the AOD-PM25 model, with a finer analysis and greater coverage.

Given that most of the current TOAR-based models are single models, this study aims to develop an integrated model using an integration technique that takes advantage of the strengths of multiple base models and combines TOAR and meteorological factor characteristics to achieve better inversion results.

2. Materials And Methods

2.1. Research Area

The Beijing-Tianjin-Hebei (BTH) region refers to the Chinese cities of Beijing, Tianjin and part of Hebei Province. The geographical range is 113.3 to 119.5°E and 36 to 42.4°N. The terrain of the Beijing-Tianjin-Hebei region is complex and varied, including plains, mountains and hills, as shown in Figure 1, left. Figure 1 right shows the Beichen District of Tianjin, which has a temperate monsoon climate with four distinct seasons. Industrial emissions, traffic exhaust and urban dust are the main sources of air pollution in Beichen District, so this study focuses on the PM2.5 concentration in Beichen District.

\*Corresponding author’s email: zhangyan@hebut.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
2.2. Data Description

The dataset mainly includes PM2.5 concentration data, satellite data and auxiliary data, as shown in Table 1. Among them, the near-surface daily mean PM2.5 station monitoring data includes data from 87 state-controlled monitoring stations in the Beijing-Tianjin-Hebei region for two years in 2019 and 2020. The satellite data come from the data measured by the MODIS sensor of Terra satellite. The L1B-class band 1 (0.62-0.67 µm) red band, band 3 blue band (0.459-0.479 µm), and band 7 (2.105-2.155 µm) short infrared band are used in this study, in addition to four angular data. Meteorological data include seven meteorological variables: boundary layer height (BLH), atmospheric surface pressure (SP), total column of water (TCW), total column of ozone (TCO), temperature at 2 m (T2M), and horizontal and vertical wind speeds at 10 m above the ground (U10M, V10M).

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable Name</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM2.5</td>
<td>PM2.5</td>
<td>-</td>
<td>day</td>
<td><a href="https://data.epmap.org/product">https://data.epmap.org/product</a></td>
</tr>
<tr>
<td>Satellite Data</td>
<td>TOAR, Solar Zenith, solar azimuth, Satellite zenith, Satellite azimuth</td>
<td>1km</td>
<td>day</td>
<td><a href="https://ladsweb.modaps.eosdis.nasa.gov/">https://ladsweb.modaps.eosdis.nasa.gov/</a></td>
</tr>
<tr>
<td>Meteorological element data</td>
<td>BLH, SP, TCW, TCO, T2M, U10, V10</td>
<td>0.25°</td>
<td>hour</td>
<td><a href="http://www.dsac.cn/">http://www.dsac.cn/</a></td>
</tr>
</tbody>
</table>

2.3. Research Methods

In the inversion study of near-surface PM2.5 concentration, the main processes in this study are data acquisition, feature screening, and inversion modelling, and the technology roadmap is shown in Figure 2.

(1) XGBoost model construction

Each time XGBoost adds a tree, it learns a new function \( f(x) \) to fit the residuals of the last prediction, and each tree will fall to a corresponding leaf node, each leaf node corresponds to a predicted value, and the score corresponding to each tree will be added up to be the predicted value of the sample. XGBoost adds a regularisation term to the objective function, which reduces the variance of the model and can effectively prevent overfitting. The objective function is shown in equation 1:

\[
Obj = \sum_{i=1}^{n} \left[ f(x_i, \hat{y}_i^{(t-1)}) + g(x_i) + \frac{1}{2} h(x_i)^2 \right] + \sum_{i} \Omega(f_i) \quad (1)
\]

Where \( \hat{y}_i^{(t-1)} \) is the model prediction value in the first t-1 rounds, \( f(x_i, \hat{y}_i^{(t-1)}) \) is the training error of sample \( X_i \).

Figure 2. Satellite inversion flow chart
\( \Omega(f_i) = \gamma T + \frac{1}{2} \|w\| \) is the regular term in XGBoost, \( T \) is the number of leaf nodes, and \( w \) is the score corresponding to each leaf node.

(2) LightGBM model construction

The LightGBM model approximates the gain of the split point by constructing a histogram of features to find the best split point. And the gradient-based one-sided sampling is integrated, so that the large gradient features can be retained while the small gradient features are randomly sampled to optimise the computation and improve the model accuracy. And in order to cope with the sparsity of high-latitude data, the mutually exclusive features are fused and bound, which ensures that the information will not be lost and reduces the amount of computation. The model objective function is as follows:

\[
Obj' = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i(x_i)) + \sum_{i=1}^{T} \Omega(f_i) \tag{2}
\]

Where \( y_i \) is the actual value of the prediction target, \( \hat{y}_i(x_i) \) is the \( t \)-th predicted value of the model, \( n \) is the number of samples, \( \Omega \) is the regular term of the model.

(3) Model integration

In model training, the results obtained by XGBoost model inversion are expressed as \( X_{\text{XGB}} = \{X_1, X_2, X_3, ..., X_n\} \). The inversion result of the LightGBM model is expressed as \( Y_{\text{LGB}} = \{Y_1, Y_2, Y_3, ..., Y_n\} \). Where \( n \) is the sequence length, the actual PM2.5 concentration data is expressed as \( Z = \{Z_1, Z_2, Z_3, ..., Z_n\} \). Translate the linear regression model \( Z = aX + bY + c \), where \( a \), \( b \), and \( c \) are the coefficients of the regression model.

(4) Evaluating Indicator

In this study, three evaluation indexes, MAE, RMSE and R2, are used to assess the model performance. MAE can demonstrate the real situation of the error of the predicted value, and the closer its value is to 0, the higher the accuracy of the model prediction results; RMSE can be used to measure the deviation between the predicted value and the real value, and when its value is closer to 0, it indicates that the model prediction results are more accurate; R2 is a statistic measuring the goodness-of-fit, and the closer its value is to 1, the better the model is fitted.

3. Results

3.1. Hyperparameter Optimisation.

In XGBoost and LightGBM, different combinations of hyperparameters lead to a large gap in the predictive performance of the models, and since both models have more hyperparameters, this study chooses the Bayesian optimisation method to optimise the main hyperparameters of the XGBoost and LightGBM models. Finally, the inversion results of the two integrated models are linearly integrated to obtain the final model fitting equation as follows:

\[
PM_{2.5, \text{inversion}} = 0.3586PM_{2.5, \text{XGB}} + 0.6815PM_{2.5, \text{LGB}} - 1.7437 \tag{3}
\]

3.2. Comparative analysis of different input features

In order to analyse the accuracy of the models, the results of AOD-based and TOAR-based models are compared and with or without the inclusion of meteorological elements.

<table>
<thead>
<tr>
<th>Name</th>
<th>Meteorological Elements</th>
<th>R2</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOAR</td>
<td>Yes</td>
<td>0.9597</td>
<td>8.5083</td>
<td>5.7101</td>
</tr>
<tr>
<td>AOD</td>
<td>Yes</td>
<td>0.9304</td>
<td>11.1838</td>
<td>7.1636</td>
</tr>
<tr>
<td>TOAR</td>
<td>No</td>
<td>0.8849</td>
<td>14.3800</td>
<td>8.7420</td>
</tr>
<tr>
<td>AOD</td>
<td>No</td>
<td>0.0918</td>
<td>40.3991</td>
<td>26.3903</td>
</tr>
</tbody>
</table>
Figure 3 shows the model performance evaluation based on different input parameters, based on TOAR on the left side and AOD on the right side, with and without meteorological element data. Table 2 shows the different model evaluation metrics. Figs. 3 (a) and (b) show the models with meteorological elements included; R2, RMSE and MAE of Fig. 3(b) are 0.9304, 11.1838 and 7.1636, respectively. Compared with Fig. 3(b), R2, RMSE and MAE of Fig. 3(a) are 0.9575, 8.7325 and 5.7181, respectively, with an improvement of 2.9% in R2, and a decrease of 2.9% in RMSE and MAE decreased by 2.67 and 1.45, respectively. This indicates that the use of TOAR instead of AOD modelling can effectively improve the accuracy of PM2.5 inversion, which is due to the fact that TOAR is a precursor product of AOD and avoids the problem of uncertainty and missing values in the AOD inversion process. Figures 3 (c) and (d) show the models without meteorological elements; comparison with the models with meteorological elements added reveals that the inversion results of the models without meteorological elements are both very unsatisfactory, but the model based on TOAR is much better than the one based on AOD, which is due to the fact that the coverage of the AOD data is extremely poor, and there are a large number of missing values within the study area, which makes it impossible to carry out the effective inversion.

3.3. Comparative analysis of different models

In order to further validate the reliability of the model proposed in this study, the more popular regression model done at this stage is selected, trained using the same dataset as in this study, and the results of the model cross-validation metrics are analysed.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>RMSE ($\mu g/m^3$)</th>
<th>MAE ($\mu g/m^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR</td>
<td>0.2369</td>
<td>35.5031</td>
<td>23.4240</td>
</tr>
<tr>
<td>GWR(^{[9]})</td>
<td>0.3604</td>
<td>22.4528</td>
<td>20.8152</td>
</tr>
<tr>
<td>RF(^{[10]})</td>
<td>0.9136</td>
<td>11.9495</td>
<td>7.0470</td>
</tr>
<tr>
<td>LightGBM</td>
<td>0.9258</td>
<td>10.1081</td>
<td>6.3903</td>
</tr>
<tr>
<td>XGBoost(^{[11]})</td>
<td>0.9133</td>
<td>11.4677</td>
<td>7.9223</td>
</tr>
<tr>
<td>RF_LGBM(^{[11]})</td>
<td>0.9404</td>
<td>10.3493</td>
<td>6.6624</td>
</tr>
<tr>
<td>This study</td>
<td>0.9574</td>
<td>8.5083</td>
<td>5.7101</td>
</tr>
</tbody>
</table>

As shown in Table 3, MLR has the worst performance, which is due to the existence of nonlinear relationships in the data, and MLR has a simple structure and the worst ability to learn nonlinearities. GWR performs slightly better than MLR, and the evaluation metrics of R2, RMSE, and MAE are optimised by 12.35%, 13.05 ($\mu g/m^3$), and 2.61 ($\mu g/m^3$), respectively, but due to the small study area, the spatial heterogeneity is not prominent, which However, due to the small study area, the spatial heterogeneity is not prominent, and the strengths of GWR in portraying spatial heterogeneity cannot be fully exploited. RF, LightGBM, and XGBoost, as the representatives of machine learning tree models, perform much better than MLR and GWR in PM2.5 concentration inversion, and the machine learning models can effectively analyse the nonlinear relationship between the data. The integrated model RF_LGBM is an integration of two base learning models, Random Forest Model and Lightweight Gradient Booster Algorithm, which combines the advantages of the two models and improves the generalisation ability and robustness of the model, and the results of the evaluation metrics, R2, RMSE, and MAE, are 0.94, 10.35 ($\mu g/m^3$), and 6.66 ($\mu g/m^3$), respectively, and the performance of the model is due to the single machine learning model. However, collectively, the XGB_LGBM model proposed in this study has the best performance due to the incorporation of regularisation and gradient descent techniques and the ability to finely tune the hyperparameters of the XGBoost model, which makes it a better predictor as a base learner compared to random forests and faster to train than random forests.

3.4. Comparative analysis of observed and predicted values

The observed values of PM2.5 concentration at Huaihe Road point of Tianjin ambient air quality monitoring station were selected and compared and analysed with the PM2.5 concentration values of this site inverted by the model of this study, and the time-varying graph is shown in Figure 4. The results show that the inversion value and the observed value in the time change trend and the value change are basically the same trend, and the model proposed in this study can well invert the PM2.5 concentration at the monitoring station.
4. Conclusions

In this study, I propose an integrated learning model based on boosting, which uses two models, LightGBM and XGBoost, as the base learner of the integrated model, and builds the integrated model by Bayesian optimisation of the hyperparameters, which has faster computing speed and better inversion.

The results show that the model adopts TOAR instead of AOD as the input parameter of the model, which achieves better inversion results. According to the model inversion results, I analysed the time series distributions of the observed and inverted values at Huaihe Station, Beichen District, Tianjin, and found that the model inversion results were highly identical to the observed values, which proved that the model could invert the PM2.5 concentration values well. My model combines a combination of punctual data sources with the advantages of continuous spatial and temporal coverage afforded by satellite remote sensing.

References