ISSN: 2664-9640

Research on $PM_{2.5}$ Retrieval Based on XGBoost

Qingxia Xue [1], Gen Zhang [1] and Xiaofang Liu [1,2]

[1] Artificial Intelligence Key Laboratory of Sichuan Province, Automation and Information Engineering, Sichuan University of Science and Engineering Zigong, 643002, China;

[2] School of Computer Science and Engineering, Sichuan University of Science and Engineering, Zigong, 643002, China.

Abstract

Most developed areas in China suffer from severe smog pollution, which may endanger public health. Satellite remote sensing data has been increasingly used to retrieve $PM_{2.5}$. Most existing studies have used polar-orbiting satellite instruments, but a major limitation of polar-orbiting platforms is that their sampling frequency is limited, which is not enough to capture $PM_{2.5}$ short-term change monitoring. The FY-4A AGRI can obtain multiple observational data in one day, which can realize the daily change monitoring of air quality, so as to study the daily change characteristics of $PM_{2.5}$. Based on this, this paper selects FY-4A AGRI L1 data combined with meteorological factors, and uses the XGBoost model to establish a ground $PM_{2.5}$ concentration inversion model. At the same time, the grid search method is used to optimize the model's hyperparameters, and an optimal $PM_{2.5}$ concentration inversion model based on XGBoost is established. This paper selects the data of the Sichuan Basin in November 2018 as the experimental data for $PM_{2.5}$ inversion and evaluation, and explores the method of using FY-4A AGRI data to invert the $PM_{2.5}$ mass concentration, and then studies the spatial distribution characteristics of air pollution in the Sichuan Basin And time variation characteristics. The results show that: based on FY-4A AGRI L1 data combined with meteorological factors, the optimized XGBoost model has the best inversion accuracy and better model performance. Among them, the MAE of the optimized XGBoost model inversion is $7.53\mu g/m^3$, the $RMSE$ is $10.7\mu g/m^3$, and the coefficient of determination is 0.84.

Keywords

XGBoost, FY-4A AGRI, $PM_{2.5}$.

1. Introduction

$PM_{2.5}$, also known as fine particulate matter, is aerosol particulate matter with a dynamic diameter $\leq 2.5\mu m$ in the air. $PM_{2.5}$ has the characteristics of small particle size, easy adsorption of toxic and harmful substances, and long transmission distance, which can not only affect the quality of the atmospheric environment, but also endanger human health [1-2]. With the rapid development of industrialization and urbanization, the environmental problems caused by $PM_{2.5}$ have become increasingly prominent. $PM_{2.5}$ has been listed as one of the main pollutants in the city and has attracted great attention from the public and the government [3-4]. Therefore, accurate prediction of $PM_{2.5}$ concentration has important practical significance for the prevention and control of air pollution and sustainable economic development. At present, China has established a $PM_{2.5}$ ground monitoring network with more than 1,600 sites. However, due to the sparse and uneven distribution of sites, it can reflect the temporal and spatial evolution of $PM_{2.5}$ in a timely and accurate manner. Satellite remote sensing has the characteristics of wide observation coverage and long-term all-weather real-time monitoring, which can supplement the deficiencies of ground monitoring sites [5-6]. Therefore, the use of

ISSN: 2664-9640

satellite remote sensing to estimate the ground $PM_{2.5}$ concentration is a current research hotspot to accurately reflect its spatial and continuous temporal and spatial changes.

The research on inversion of $PM_{2.5}$ is mainly based on the physical relationship between atmospheric extinction parameters such as AOD (Aerosol Optical Depth) and the ground $PM_{2.5}$ concentration, combined with meteorological conditions, aerosol characteristics and other auxiliary parameters to establish a physical model to estimate $PM_{2.5}$ concentration [7-8]. The exponential increase in physical models and data requires more computing resources and time, which is not conducive to global or large-scale inversion. In addition, the temporal and spatial distribution of $PM_{2.5}$ concentration is affected by various factors such as meteorological fields, emission sources, complex underlying surfaces, and the coupling of physical, chemical and biological processes in the atmosphere, and has strong nonlinear characteristics. And machine learning can extract effective information from massive data under the condition of insufficient understanding of physical mechanism, making it an important method for the development of satellite remote sensing inversion.

Therefore, some scholars try to use satellite remote sensing AOD, meteorological conditions obtained from model analysis fields as input, and use multiple linear regression and support vector machine methods to replace complex physical model construction. The MODIS data [9-10] obtained better results. Some scholars [11] used neural network methods to build more complex implicit association models. However, neural networks have shortcomings such as poor generalization ability, over-fitting, and complexity in finding structural parameters [6]. In response to this problem, some scholars have introduced the Random Forest [12-13] method for inversion, which has been well applied. Subsequently, Pan [14] applied the XGBoost (Extreme Gradient Boosting) algorithm to predict the $PM_{2.5}$ hourly concentration in China. The results of random forest, support vector machine, multiple linear regression and decision tree regression were compared, and the results showed that the performance of the XGBoost algorithm is better than other data mining method.

Retrieving $PM_{2.5}$ is currently mostly based on MODIS data, but the sensors represented by MODIS have the problem of too low time and space resolution in acquiring satellite data. Satellite observation data for 1 to 2 times a day cannot achieve intra-day change monitoring of air quality [15]. The FY-4A (Fengyun-4a) is China's second-generation geostationary orbital meteorological satellite. Its AGRI (Advanced Geosynchronous Radiation Imager) adopts the working method of full-disk scanning imaging, which can obtain multiple images in one day. Observation data can effectively make up for the lack of space and time coverage of ground $PM_{2.5}$ monitoring sites.

This paper uses FY-4A AGRI L1 data to construct an XGBoost model to invert and evaluate $PM_{2.5}$ in the Sichuan Basin in the fall of 2018, explore the method of using FY-4A AGRI data to retrieve $PM_{2.5}$ mass concentration, and then study Sichuan the spatial distribution and temporal variation characteristics of air pollution in the basin.

2. Research methods

2.1. Spatio-temporal matching

Match each satellite pixel and meteorological data pixel to the ground information data with the closest pixel point, as shown in equation (1):

$$d_{\min} = \sqrt{\left(lon - lon_{gro}\right)^2 + \left(lat - lat_{gro}\right)^2} \tag{1}$$

Among them, $d_{\min}$ represents the minimum distance between two points; lon represents the longitude value of a certain pixel of satellite or meteorological data; lon_{gro} represents the longitude value of a certain pixel of a ground station; lat represents the latitude value of a

ISSN: 2664-9640

certain pixel of satellite or meteorological data; lat_{gro} represents the longitude value of a pixel on the ground station.

2.2. XGBoost Model

XGBoost is an extreme gradient boosting algorithm and an integrated tree Model, the objective function of XGBoost is:

$$\begin{cases} obj = \sum_i l(\hat{y}_i + y_i) + \sum_k \Omega(f_k) \\ \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \end{cases} \tag{2}$$

Among them, *obj* is the loss function, used to measure the difference between the predicted value $\hat{y}_i$ and the true value y_i, f_k represents the *k-th* tree model, and the second term $\Omega(f_k)$ is the penalty function, that is, the complexity of the penalty model. In the penalty function, γ is the complexity parameter, *T* is the leaf node tree, and λ is the penalty coefficient of the leaf weight ω. The penalty function $\Omega(f_k)$ helps to smooth the final learned weights to avoid overfitting. The XGBoost model is a stepwise additive model, which is formed by adding *k* tree models. The complete iterative decision tree is:

$$\hat{y}_i^{(k+1)} = \hat{y}_i^{(k)} + \eta f_{k+1}(X_i) \tag{3}$$

f_{k+1} is the *k+1* tree model, η is the step length of the iteration, that is, the learning rate, X_i represents the *i-th* instance. The size of η determines the iteration speed. The smaller the η the slower the convergence speed, but a more accurate optimal value can be found.

Let $\hat{y}_i^{(k)}$ be the predicted value of the *i-th* instance X_i in the *k-th* iteration, use $\hat{y}_i$ in equation (2) as a parameter, and add f_k to minimize the objective function:

$$obj^k \approx \sum_{i=1}^{n}[g_i f_k(X_i) + \frac{1}{2} h_i f_k^2(X_i)] + \sum_{i=1}^{k}\Omega(f_i) \tag{4}$$

Among them, g_i and h_i are the first step statistics and the second step statistics of the loss function.

Define $I_j = \{i \mid q(X_i) = j\}$ as the sample number set of the *j-th* leaf node, and rewrite formula (4) as:

$$obj^{(k)} = \sum_{j=1}^{T}[(\sum_{i\in I_i} g_j)\omega_j + \frac{1}{2}(\sum_{i\in I_i} h_j + \lambda)\omega_j^2] + \gamma T \tag{5}$$

Obtain the first-order partial derivative of the objective function of formula (6) with respect to ω_j, and set it equal to 0 to obtain the optimal weight corresponding to the leaf node *j*:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_j}{\sum_{i\in I_j} h_j + \lambda} \tag{7}$$

And the optimal value of the objective function is:

$$obj^{\mathrm{k}} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_j + \lambda} + \gamma T \tag{8}$$

ISSN: 2664-9640

Equation (7) is also a structure score, which is used to judge the quality of the structure tree. It is obtained for different loss functions.

XGBoost performs a second-order Taylor expansion on the cost function, and uses the first-order and second-order derivatives at the same time, which improves the accuracy of the model. XGBooost adds a regular term to the loss function to control the complexity of the model. Its regular term reduces the variance of the model, makes the learned model simpler, and can effectively prevent the model from overfitting [16].

2.3. Evaluation Index

This article uses common evaluation indicators RMSE, MAE and R^2 to compare model accuracy. The calculation formula of the indicator is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2} \tag{11}$$

y_i is the true value of the *i-th* sample; $\hat{y}_i$ is the predicted value of the *i-th* sample; $\bar{\hat{y}}$ is the average of the predicted values; n is the sample size. MAE reflects the actual situation of the predicted value error, $RMSE$ reflects the degree of deviation between the predicted value and the true value, the smaller the value of the two. R^2 reflects the degree of fit between the predicted value of the model and the actual value. The closer the value is to 1, the better the effect of the model.

3. Experiment and analysis

3.1. Research area and data

3.1.1. Research area

The Sichuan Basin is located in southwestern China, with a total area of more than 260,000 square kilometers (28° N～32°40'N, 103°11'N～107°47'N) (see Figure 1). The cooling effect of the edge mountains on the air causes the cold air to move along the bottom of the basin, which is easy to accumulate at the bottom of the basin, forming a typical terrain inversion phenomenon, causing continuous accumulation of pollutants and deteriorating regional air quality [17].

3.1.2. Experimental data

(1) Remote sensing data source: This paper uses FY-4A AGRI full disk space 4 km resolution L1 level data, using 14 channels of data to retrieve $PM_{2.5}$. AGRI has 14 channels, covering from 0.47 microns (visible light band) to 13.5 microns (thermal infrared band), of which 6 are visible/near-infrared bands, 2 are mid-wave infrared bands, 2 water vapor bands and 4 long-wave infrared bands Band [18].

(2) Meteorological data: The meteorological reanalysis data selected in this paper are ERA5 data published by ECMWF[19]. Using ECMWF-ERA5 weather model data with a time resolution of hour by hour, a spatial resolution of 975hPa, and a pressure level of 975hPa in November of 2018, the meteorological variables include relative humidity (RH), temperature (Tem), and meridional Wind speed (U), zonal wind speed (V) and terrain (Geo).

ISSN: 2664-9640

(3) Ground $PM_{2.5}$ data: This article uses ground $PM_{2.5}$ data from the Ministry of Ecology and Environment of the People's Republic of China (http://106.37.208.233:20035/), of which 108 sites are located in the Sichuan Basin. The site distribution is shown in Figure 1, using hourly data for model construction.

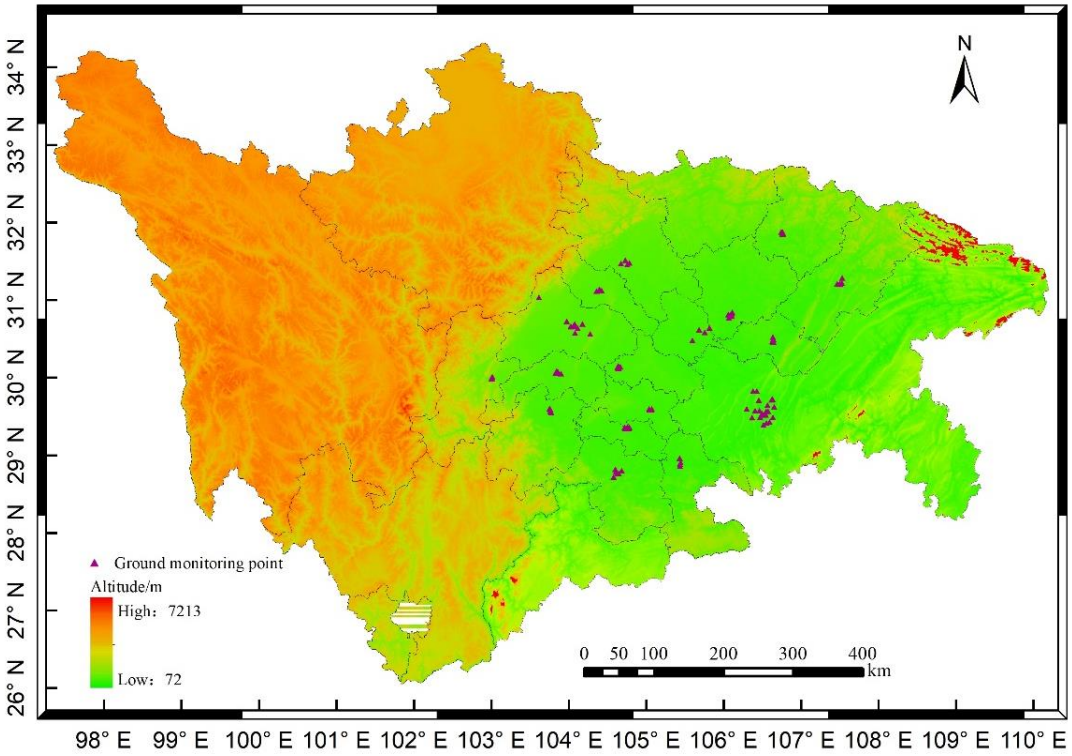

Figure 1: The geographical location of the study area and the distribution of $PM_{2.5}$ ground monitoring stations

3.2. Experiment content

The flowchart of XGBoost model inversion of ground $PM_{2.5}$ concentration is shown in Figure 2.

3.2.1. Building dataset

The data used is the hourly data of the Sichuan Basin in November 2018 (UTC: 2018-11-01 00:00:00~2018-11-30 23:00:00), and the calibrated FY-4A AGRI L1 data (including 14 channels), meteorological data 975hPa atmospheric pressure meteorological data-RH, Tem, U, V, Geo and station latitude and longitude (lon, lat), $PM_{2.5}$ concentration hourly value ($PM_{2.5}$_Hourly) according to formula (1) Perform spatio-temporal matching.

After removing the missing values of satellite data, the zero values of visible light and near-infrared channel data, and the missing values of $PM_{2.5}$ hourly concentration at ground stations in the obtained data set, a total of 31,194 data items were removed. 80% of the data is used for training and 20% of the data is used for testing.

3.2.2. Model construction and parameter adjustment

This paper selects the following four hyperparameters of XGBoost for tuning: the number of trees n_estimators, the maximum depth of the tree max_depth, the minimum separation loss value gamma, and the smallest sample weight and min_child_weight among the child nodes.

First, a random search is performed on XGBoost hyperparameters through multiple 5-fold cross-validation to obtain multiple optimal hyperparameter combinations, see Table 1.

ISSN: 2664-9640

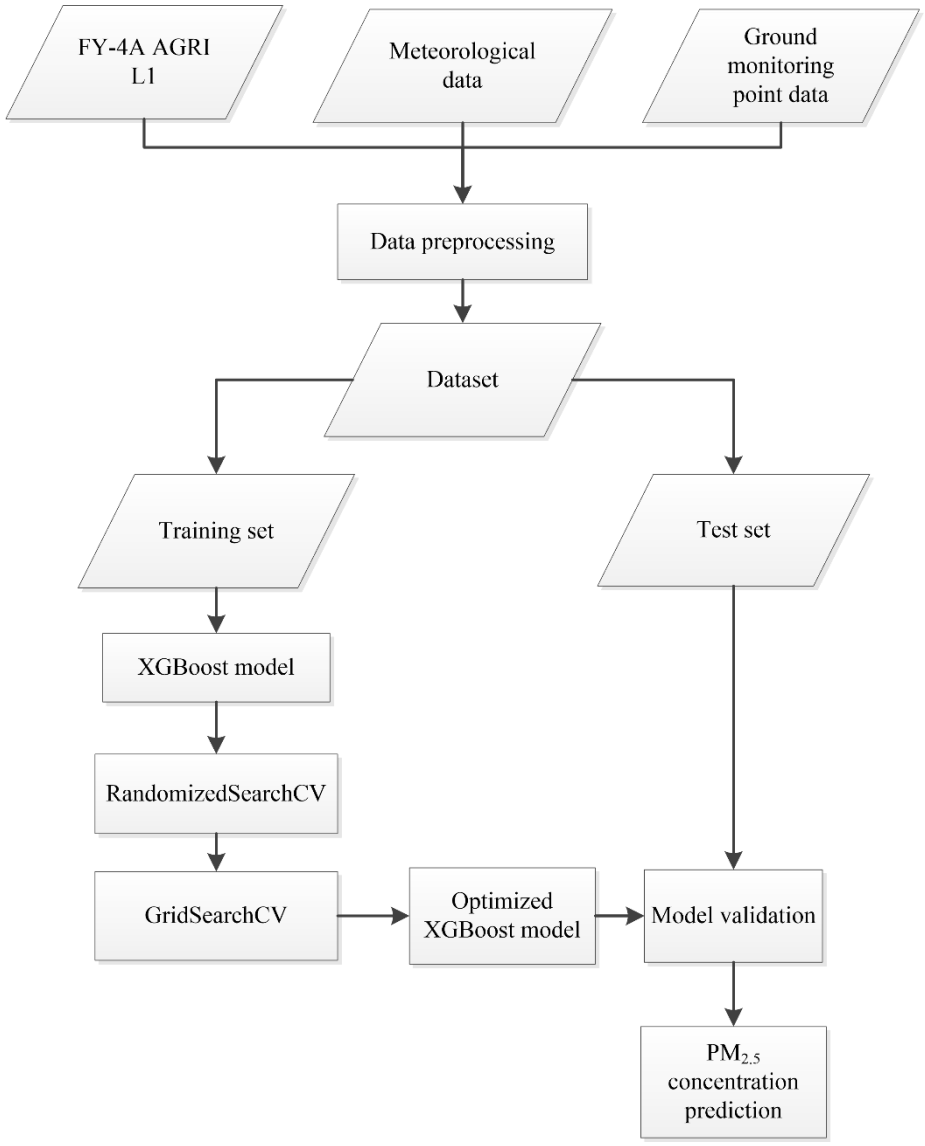

Figure 2: The flowchart of XGBoost model inversion of ground $PM_{2.5}$ concentration

Table 1: Basic situation of XGBoost random search hyperparameters

Parameter Name	Adjustment Range	output value
n_estimators	(50,3000)	200,340,150
max_depth	(1,50)	11,16,13
gamma	(0,1)	0.7,0.7,0.8
min_child_weight	(1,10)	8,7,9

Then, on the basis of the optimal results of random matching, the range of hyperparameters is reduced, and each match is traversed through grid search, so as to obtain the final value of hyperparameters. The basic situation of XGBoost grid search hyperparameters see Table 2.

Table 2: The basic situation of XGBoost grid search hyperparameters

Parameter Name	Adjustment Range	Best Value
n_estimators	(100,350)	180
max_depth	(10,18)	15

ISSN: 2664-9640

gamma	[0.7,0.8]	0.7
min_child_weight	(7,10)	8

3.3. Experimental results and analysis

3.3.1. Model construction and parameter adjustment

Use the saved XGBoost model to analyze and evaluate the importance of input variables, and the result of variable importance is shown in Figure 3. The order of importance of the variables used in this article is temperature, relative humidity, geopotential, Channel10, Channel03, zonal wind, meridional wind, longitude, Channel09, latitude, Channel13, Channel06, Channel04, Channel14, Channel11, Channel01, Channel12, Channel08, Channel07, Channel02, Channel05. In FY 4A AGRI L1 variables, the importance of Channel10, Channel03 and Channel10 is relatively high, while Channel02 and Channel05 are relatively low. Among the meteorological factors, the importance of temperature and relative humidity is relatively high, and the zonal wind and meridional wind are relatively low.

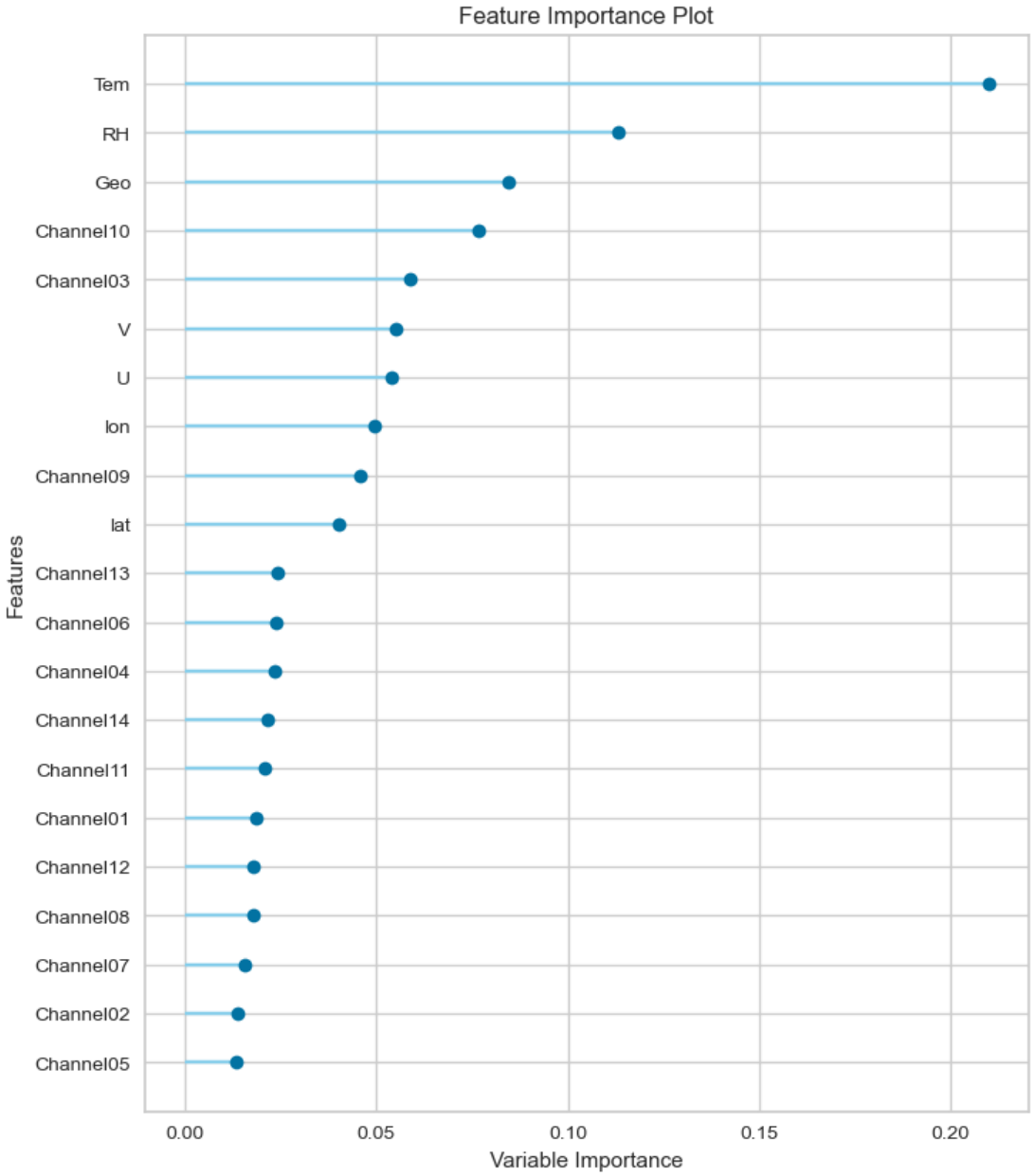

Figure 3: Variable importance

ISSN: 2664-9640

3.3.2. Model construction and parameter adjustment

The inversion accuracy of the model is mainly used to evaluate the prediction effect of the test set by the model, and the evaluation indicators use MAE, RMSE and. Figure 4 shows the ground $PM_{2.5}$ inversion results of each algorithm. It can be seen from Figure 4 that compared to decision trees and knns that do not use integrated strategies, using Bagging integrated Random Forest and Boosting integrated XGBoost respectively, the performance indicators of the model have been greatly improved. It can be seen that the integrated strategy has better performance than the result of the single learner model.At the same time, the XGBoost model performs better than the random forest. Among them, the MAE of the XGBoost model is $7.53\mu g/m^3$, the $RMSE$ is $10.7\mu g/m^3$, the R^2 is 0.84, and 63.71% of the inversion results are within the expected error range. The fitting slopes of the four models are all less than 1, indicating that the models all have a certain degree of deviation. But comparing the three models, the XGBoost model has the highest slope, reaching 0.83. It shows that the XGBoost model has the best fitting effect.

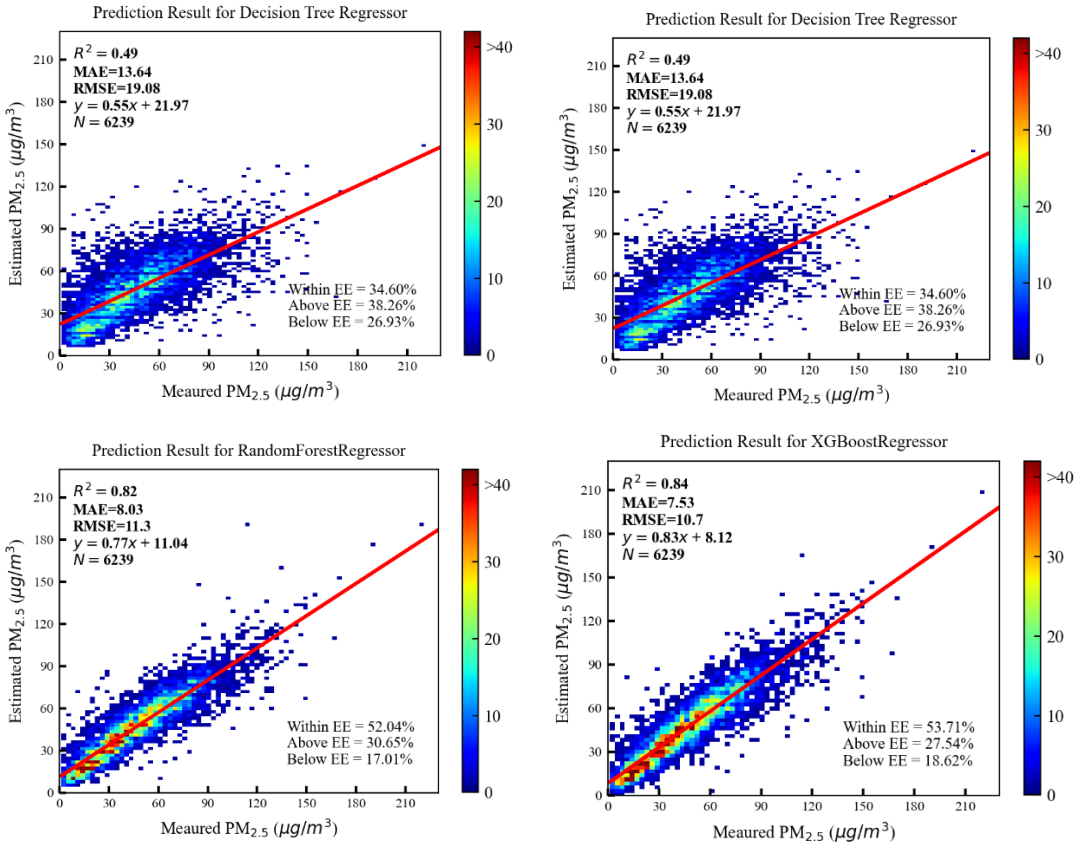

Figure 4: Ground $PM_{2.5}$ inversion results of each algorithm

3.3.3. Analysis of PM2.5 concentration in the Sichuan Basin in November 2018

Figure 5 is a columnar comparison between the monthly mean value of $PM_{2.5}$ retrieved by the XGBoost model in the Sichuan Basin and the monthly mean value of $PM_{2.5}$ at the ground station. It can be seen from Figure 5 that the two mean values are not much different. It shows that the spatial distribution of the two is similar, and the model inversion result has certain reliability. Among them, Leshan City and Chongqing City have the best fit, and the two cities with the highest $PM_{2.5}$ concentration values in November 2018 are Zigong City and Yibin City.

ISSN: 2664-9640

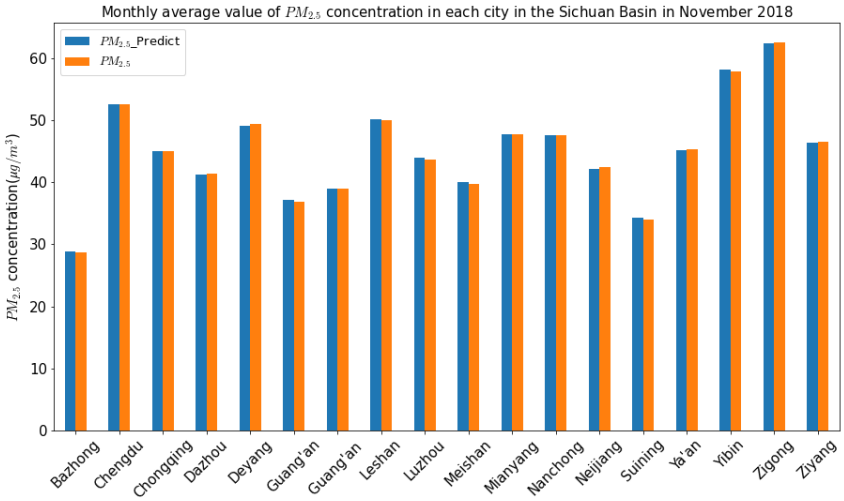

Figure 5: The histogram of the monthly mean value of $PM_{2.5}$ inverted by the XGBoost model in the Sichuan Basin and the monthly mean value of $PM_{2.5}$ at ground stations

4. Conclusion

In this paper, the AGRI L1 level data carried on the FY-4A geostationary satellite is used as the remote sensing data source, combined with meteorological factors to construct the optimal hyperparameter XGBoost method to retrieve the ground $PM_{2.5}$ concentration value. The results show:

(1) The analysis of the importance of model variables shows that the importance of temperature and relative humidity in meteorological factors is relatively high, indicating that temperature and relative humidity have a greater impact on $PM_{2.5}$. The importance of Channel10 and Channel03 in FY-4A AGRI L1 data is relatively high. Among them, the channel type of Channel10 is water vapor, and the channel type of Channel03 is visible light and near-infrared with a spectral bandwidth of 0.75～0.90μm. This means that the water vapor and visible light and near-infrared with a spectral bandwidth of 0.75～0.90μm are FY-4A AGRI L1. Data inversion is one of the important factors of $PM_{2.5}$.

(2) The XGBoost method established based on FY-4A AGRI data combined with auxiliary meteorological factors can better retrieve ground $PM_{2.5}$. Then use random search and then grid search to optimize the hyperparameter tree of XGBoost, which further improves the prediction accuracy. This paper uses the hourly data of the Sichuan Basin in November 2018 as the experimental data, and uses the method proposed in this paper for model training and testing, which verifies that the XGBoost model has good performance in retrieving ground $PM_{2.5}$ concentration with high temporal and spatial resolution. .

(3) There are still shortcomings in this article. Since the time range of the data set used in this article is only one month, the seasonal characteristics of $PM_{2.5}$ have not been studied yet.

Acknowledgements

This work is supported in part by the Science and Technology Department Project of Sichuan Provincial of China, under Grant 2017GZ0303, in part by Academician (Expert) Workstation Fund Project of Sichuan Province of China, under Grant 2016YSGZZ01, in part by Special Fund for Training High Level Innovative Talents of Sichuan University of Science and Engineering, under Grant B12402005, and Sichuan University of Science and Engineering for Talent introduction project, under Grant 2021RC16.

ISSN: 2664-9640

References

[1] M. Miri, A. Alahabadi and M.H. Ehrampush: Mortality and morbidity due to exposure to ambient particulate matter, Ecotoxicology and Environmental Safety, Vol. 165 (2018) No.19, p.307-313.

[2] A.J. Cohen Author, M. Brauer and R. Burnett: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The Lancet, 2017, Vol. 389 (2017) No. 10082, p. 1907-1918.

[3] L. Sun, J. Wei and D.H. Duan: Impact of Land-Use and Land-Cover Change on urban air quality in representative cities of China, Journal of Atmospheric and Solar-Terrestrial Physics, Vol. 142 (2016) No.6, p.43-54.

[4] J. Wei, W. Huang and Z.Q. Li: Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach, Remote Sensing of Environment, Vol. 231 (2019) No.12, p.111221.

[5] X. Yu, W.J. Zhao and C.Y. Sun: Progress study on remote sensing retrieval of atmospheric PM2.5 concentration, Environmental Pollution & Control, Vol. 39 (2017) No.10, p.1153-1158.

[6] M.X. Huang, B. Wei and Q.T, Hao: review on research of PM2.5 retrieval by remote sensing technology, Environmental Pollution & Control, Vol. 37 (2015) No.10, p.70-76+85.

[7] T.W. Li, Y.Q. Sun and C.X. Yang: Retrieving PM2.5 Using Satellite Remote Sensing and Ground Station Measurements, Journal of Geomatics, Vol. 40 (2015) No.03, p.6-9.

[8] Q. Jiang, C. Ying and F. Wang: Estimates and verification of surface PM2.5 mass concentration in China based on FY-4A satellite optical products, Acta Meteorological Sinica, Vol. 79 (2021) No.03, p.492-508.

[9] Y.L. Sun, X.J. Cui and J.N. Xiong: Junnan:Inversion and Analysis of PM2.5 Concentration in Chengdu by Remote Sensing, Journal of Geomatics, Vol. 46 (2021) No.S1, p.75-81.

[10] H. Zhang, S.G. Wang and J.Y. Xin: The temporal and spatial distribution characteristics of PM2.5 in the Sichuan Basin based on MODIS AOD revised by ground-based observations, Journal of Lanzhou University (Natural Sciences), Vol. 55 (2019) No.05, p.610-615+623.

[11] Y.R. Wu, J.P. Guo and X. Zhang: Correlation between PM concentrations and Aerosol Optical Depth in eastern China based on BP neural networks, IEEE International Geoscience and Remote Sensing Symposium, (Vancouver, BC, Canada, July 24-29, 2011). Vol. 24, p.3308-3311.

[12] Q. Shao, Y.H. Chen and J. Li: Inversion of PM2.5 Concentration in Beijing Based on Satellite Remote Sensing and Meteorological Reanalysis Data, Geography and Geo-Information Science, Vol. 34 (2018) No.03, p.32-38.

[13] X. Du, J.Y. Feng and S.Q. Lv:PM2.5 concentration prediction model based on random forest regression analysis, Telecommunications Science, Vol. 33 (2017) No.07, p.66-75.

[14] B.Y. Pan: Application of XGBoost algorithm in hourly PM2.5 concentration prediction, IOP Conference Series: Earth and Environmental Science, Vol. 113 (2018) No.01, p.012127.

[15] Z.P. Li, J. Chen: Remote Sensing Retrieval of Atmospheric Fine Particle PM2.5 based on GOCI Satellite and Its Temporal and Spatial Distribution, Remote Sensing Technology and Application, Vol. 35 (2020) No.01, p.163-173.

[16] J.F. Kang, J.L. Tan and L. Fang: Short-term PM2.5 concentration prediction based on XGBoost and LSTM variable weight combination model: a case study of Shanghai, China Environmental Science, Vol. 41 (2021) No.09, p.4016-4025.

[17] Y.L. Tang, F.M. Yang and Y. Zhan: High resolution spatiotemporal distribution and correlation analysis of PM2.5 and PM10 concentrations in the Sichuan Basin, China Environmental Science, Vol. 39 (2019) No.12, p.4950-4958.

[18] Information on: http://gsics.nsmc.org.cn/portal/cn/instrument/AGRI.html.

[19] Information on: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels? tab= overvi ew.