Hindawi
Journal of Sensors
Volume 2022, Article ID 7148682, 8 pages
https://doi.org/10.1155/2022/7148682

Research Article

Enhancing the Applicability of Satellite Remote Sensing for $PM_{2.5}$ Estimation Using Machine Learning Models in China

Jun Chai,[1,2] Jun Song,[3] Yawen Xu,[2] Le Zhang,[1] and Bing Guo[1]

[1]*College of Computer Science, Sichuan University, China*
[2]*Chengdu Hankang Information Industry Co., LTD, China*
[3]*Department of Geography, Faculty of Social Sciences, Hong Kong Baptist University, China*

Correspondence should be addressed to Jun Song; j.song17@imperial.ac.uk and Bing Guo; guobing@scu.edu.cn

Received 20 August 2022; Accepted 16 September 2022; Published 27 September 2022

Academic Editor: Yuan Li

Numerous studies and monitoring data indicate that fine particle ($PM_{2.5}$) pollution in China is still comparatively severe. Given the sparse and uneven distribution of air quality monitoring base stations established in China and the limitation of geographical conditions, inversion of aerosol optical depth by satellite remote sensing can achieve low-cost air quality monitoring in global areas. In this study, we use the machine learning algorithm XGBoost to build a prediction model to achieve nationwide average $PM_{2.5}$ concentration prediction. Meanwhile, we used aerosol data from Moderate Resolution Imaging Spectroradiometer (MODIS) in a specific band, combined with a land use regression (LUR) model as predictors of surface $PM_{2.5}$ concentrations in China, for the period Dec. 2019-Nov. 2021. In order to provide more accurate $PM_{2.5}$ concentration prediction, the correspondence between $PM_{2.5}$ and aerosol optical depth (AOD) under different seasons was studied. The coefficients of determination (R2) for different seasons are 0.86 (spring), 0.80 (summer), 0.90 (autumn), and 0.88 (winter), indicating that the fit is best for autumn and worse for summer. The study shows the potential usefulness of using the LUR model with the XGBoost algorithm for predictive assessment of $PM_{2.5}$ spatial distribution.

1. Introduction

With the accelerated development of domestic industrialization, air pollution has progressively escalated, and relevant studies have confirmed that the increase in the concentration of fine particulate pollutants in the atmosphere is closely related to the increase of mortality. The exposure response model indicated that when $PM_{2.5}$ concentration increased by 100 $\mu g/m^3$, respiratory diseases, cardiovascular diseases, coronary heart disease, stroke, and chronic obstructive pulmonary disease were 8.32%, 6.18%, 8.32%, 5.13%, and 7.25%, respectively. Air pollution has a long-term and short-term impact on disease death. [1, 2]. $PM_{2.5}$, an important component of haze, can induce or aggravate diseases of various systems [3]. Compared with larger particles, fine particles less than 2.5 microns are more likely to enter the human body's gas exchange and blood circulation system, which not only destroys the ventilation function of regional bronchi and alveoli but also causes inflammation, leading to dysfunction of blood vessels and cells [4, 5]. In order to monitor changes in particulate matter concentrations ($PM_{2.5}$) nationwide, China has established air quality monitoring stations covering major cities and regions since January 2013. However, the monitoring stations are sparsely distributed, and most of them are located in urban areas. Further, current monitoring site data are not suitable for regional $PM_{2.5}$ concentration change studies and studies focusing on rural areas, etc.

Air quality monitoring stations are sparsely distributed, and there is a lack of monitoring means in some regions with harsh geographical conditions. Nevertheless, the method of predicting the concentration of air pollutants by inversion of aerosol by satellite remote sensing has the advantage of high efficiency and low cost [6–12]. Because remote sensing satellite coverage is prevalent, large area synchronized observation can be done in a short period of time, handy and quick access to real-time global range of all kinds of natural phenomenon is the latest information, so using the satellite

remote sensing data information inference method is convenient, and efficiency of the $PM_{2.5}$ is artificial measurement, and ground base station monitoring is unable to compare [13, 14]. The high spatial resolution of air pollutant concentration anticipated by remote sensing satellite inversion is beneficial to evaluate the air pollution index and to carry out epidemiological research on air pollutant exposure. Although satellite images can provide AOD data covering the Earth's surface, these remotely sensed images are susceptible to cloudy weather and water/snow glow reflections [15, 16].

There are a number of statistical models that link pollutant concentrations to AOD [17–22], among which the land use regression (LUR) model can accurately quantify spatial and temporal trends of pollutants at small scales [23]. The LUR model is an efficient method to evaluate the spatial variation of pollutants. It utilizes monitoring station data combined with multiple parameters such as land use, traffic information, and population density distribution to forecast and evaluate the pollutant concentration in areas not covered by monitoring station distribution through a statistical regression method. The pollutant concentration prediction model was constructed by selecting the source of fine particle sample data and the characteristic variables of land use data [24–26]. A variety of LUR models with different temporal resolutions are currently being developed in China using various technologies. The LUR model employs statistical pollutant source data after correlation analysis as the forecast dependent variable and accurate multivariate data sets of natural geographical conditions such as land use type, terrain distribution, and natural climate type as the prediction independent variable and selects data from 20-100 monitoring stations to establish multiple linear regression mapping [27, 28]. Now, some research for different environment LUR model made a lot of extensibility of development, for example, according to the different characteristics of seasonal change to develop and adapt to climate change caused by a seasonal temperature LUR model [29, 30]; other studies focusing on the time change tendency of air pollutant concentration change trend prediction question have carried on the research and development. The change trend and evaluation value of pollutants in the next few weeks are predicted through historical measurement data [31].

In this study, we used the ML-based LUR model to estimate the daily ground-level $PM_{2.5}$ concentrations in China for the period December 2019 to November 2021. We used the MODIS remote sensing satellite data product, which carries the Moderate Resolution Imaging Spectroradiometer (MODIS) on Terra and Aqua, an important instrument for observing biological and physical processes around the globe. MCD19A2 V6 is an AOD raster data product that can achieve multiangle correction. It is a level 2 data product that is generally used after calibration and positioning processing, and the raster resolution is 1 km. We used a new feature engineering approach to construct a high-resolution grid mapping by combining air pollutant concentrations, land use, meteorological factors, and AOD data as predictors of the model with the advanced machine learning algorithm XGBoost to characterize the spatial and temporal evolution of $PM_{2.5}$ concentrations at the national scale. To ensure the accuracy of the data, the experiment uses observations from over 2400 national weather stations, a sample of nearly 1600 ambient air quality monitoring sites, and satellite-based AOD retrieval data to train the model. The results of the study will help to enhance the analysis of the near-ground $PM_{2.5}$ pollution situation and the understanding of the spatial and temporal evolution of $PM_{2.5}$ pollution in China by policy-makers.

2. Materials and Measurements

2.1. Ground-Level AQ Measurements. Daily hourly $PM_{2.5}$ measurements near the ground in China were obtained from the China National Environmental Monitoring Station (CNEMC, http://www.cnemc.cn, December 1, 2021). The data provided by the website are hourly, corresponding to different detection points in each city, and contain six reference indicators: O_2, NO_2, PM_{10}, $PM_{2.5}$, O_3, and CO, of which we obtain $PM_{2.5}$ data samples; meanwhile, to ensure the accuracy of the data, only monitoring data from government environmental monitoring stations are used in this paper. The data from monitoring stations may have extreme values and missing values due to machine failure, bad weather, etc. Therefore, we need to screen the abnormal values and make up the sliding window for the vacant values before using them to ensure the continuity and validity of the model input monitoring data.

2.2. Satellite AOD Data. Aerosol data are derived from MODIS remote sensing satellite products, among which the MCD19A2 scientific data set provides products including MAIAC atmospheric correction multidimensional reflectivity band data. The orbit with the largest coverage is selected for processing according to the number of satellite transits. Access is to 1 km resolution secondary data products from Terra and Aqua satellites. The raw data included in this product are mainly AOD at 0.47 μm and 0.55 μm, AOD uncertainty at 0.47 μm range 0 to range 4, fine mode fraction for ocean, column water vapor in cm liquid water, regional background model used, cosine of solar zenith angle, relative azimuth angle, etc. The purpose of atmospheric correction (MAIAC) [32] is to eliminate the influence of atmosphere, light, and other factors on ground object reflection, so as to retrieve real physical model parameters such as reflectance, radiation rate, and surface temperature. Incorporating multiple wavelengths of the sun and zenith angle and azimuth angle parameters of the satellite information, through the radiative transfer model inversion algorithm band operation, remove reflecting solar, sensor, and the target value for atmospheric path length difference between the impact of different regions of elimination, different objects, and different light and shade as yuan after the grey value influence of the aerosol optical thickness parameter. This study focuses on the analysis of AOD data measured at 0.47 μm wavelength between December 2019 and November 2021, with the improved MAIAC product (MCD19A2) having a better spatial resolution, and the data

set is a daily product, and the data processing is still tedious when the study area is national and the time series is long.

As shown in Figure 1, the AOD data from the MCD19A2 version 6 aerosol product can reflect the spatial distribution of aerosol levels well. The data are selected from the blue-band AOD at 0.47 μm wavelength, but the directly obtained orbital aerosol data are not enough to cover the whole country, so the HDF4 data of 22 orbits need to be converted to TIF format by the MRT (MODIS Reprojection Tool) provided by NASA and then stitched; meanwhile, this study takes 1 km × 1 km. The AOD value at the center of spatial resolution is used as a representative estimate for each 25 km × 25 km meteorological grid center for subsequent model analysis.

2.3. Meteorological Information. The data set used in this paper is from the National Weather Administration of the United States Weather Service. The Climate Prediction Center is responsible for providing short-term weather fluctuation monitoring and forecasting and long-term climate change impact studies. The data sets using the advanced global data assimilation system will interpolate observation data and instrument monitoring data to a three-dimensional grid. The grid provides the forecast of output data, combined with the improvement of the global telecommunications system database and other monitoring station sources of statistical data collection, analysis, quality control, and assimilation process after finishing to obtain a complete set of data. The data set is selected for the time range from December 2019 to November 2021, and the gridded meteorological parameters mainly include temperature, relative humidity, pressure, and wind speed. The grid size is selected to be 25 km × 25 km. meteorological information is obtained every six hours, and the daily data averages of 0, 6, 12, and 18 o'clock are selected as the daily meteorological information in this study to achieve a day-based meteorological data set produced in days.

2.4. Land Use Predictors. The LUR model is used to extract the relevant factors affecting $PM_{2.5}$ concentration based on the GIS platform (ArcGIS 10.8), such as meteorology, topography, and land use. The spatial distribution of near-surface $PM_{2.5}$ concentration in China is predicted and analyzed by combining the national $PM_{2.5}$ concentration ground monitoring data, and the reasons affecting the prediction accuracy are explored in order to provide a database for the study of air quality and its impact on human health.

The following variables were considered as predictive factors:

Air pollutants: pollutants that cause harmful effects on the human living environment selected, such as $PM_{2.5}$ and PM_{10}, were at near-ground concentrations. The data set is based on the pollutant data collected by CNEMC from air quality monitoring stations.

Meteorological factors: data from the NCEP sites, including the United States national weather bureau issuing a series of weather-related business data, covers the global meteorological monitoring site detailed record of the daily and hourly weather data measured records, high-resolution satellite data, the environmental monitoring data, and other fields. We extracted meteorological data such as air relative humidity, climate temperature, and wind levels.

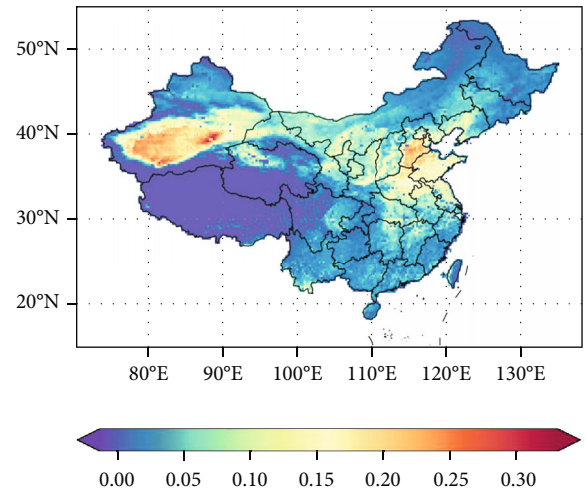

Figure 1: Spatial distribution of AOD (average value in time slice window) for the period of December 2019-November 2021.

Land use factors: the concentration of air pollutants is highly related to certain land use types; for example, forest land and green land can reduce some air pollution, while urban planning land and industrial land generally have a higher pollution index. Using remote sensing data, the cumulative area of various land types within different stations is calculated as the independent variable of land use types.

Compared with satellite AOD data, these predictive factors can reflect the influence of local sources on $PM_{2.5}$ concentrations at a more accurate spatial resolution.

2.5. Feature Engineering Approach. The feature classification involved in the air quality model is shown in Table 1, which is generally divided into dynamic and static categories. Static features include land use types, AOD data, latitude and longitude information, time information, and digital elevation; dynamic features refer to meteorological parameters obtained from meteorological data sets, three of which are selected in this study: temperature, relative humidity, and wind speed. In order to construct an optimal model for long-term prediction of air pollutants, we adopt a new feature engineering approach in order to reduce time consumption with guaranteed accuracy. In this study, 1/3 of the training data from the overall sample data set is taken out, and the importance of all features is obtained and ranked from highest to lowest by training the model, and finally, the top 30 features are used for model fitting, validation, and analysis.

2.6. Development of ML-LUR Model. XGBoost (eXtreme Gradient Boosting) [33] is an algorithm model framework based on the lifting tree, which is very powerful in distributed parallel computing efficiency, missing value processing, and prediction performance.

In this study, we compare it with other surrogate models integrating with LUR, including the standard land use

Table 1: Supporting features used in this study.

	Data category	Data type
Static	Geographic and land use	Area of wood land cover (km^2)
		Area of grass land cover (km^2)
		Area of construction land (km^2)
		Area of ocean (km^2)
		POIs (19 categories)
		AOIs (19 categories)
		Digital elevation (m)
Dynamic	Meteorology	RH (%)
		Temperature (°C)
		Wind speed (m/s)
	Remote sensing	AOD retrievals

regression (LUR) [34], *K*-Nearest Neighbors (KNN) [35], Auto Encoder (AE) [36], Support Vector Regression (SVR) [37], Deep Air Learning (DAL) [38], and Gaussian Process Regression (GPR) [39]. We evaluate the model performance by 10-fold cross-validation (CV) tests in consideration of their prediction accuracy and robustness.

Compared to others, XGBoost [40] has faster training and less memory usage, handles category features, greatly speeds up training, and has better accuracy. Therefore, we constructed ML-LUR to use XGBoost as a surrogate model for robust space-time estimation of $PM_{2.5}$ concentrations.

2.7. Model Validation. We used the LUR model based on machine learning to explain the impact of characteristic parameters combined with land use information and meteorological conditions on $PM_{2.5}$. The research content carried out daily concentration calibration according to the influence of seasonal change and yearly change trend of $PM_{2.5}$, so as to gain an accurate prediction of $PM_{2.5}$ daily concentration. First, the top 30 features were selected to build the feature data set based on the importance ranking of the features, according to the solar calendar. The study selected two years from December 2019 to November 2021 to construct the pollutant fitting model. In our experimental evaluation, 10-fold cross-validation was used to decrease contingency, and multiple partitions of the data set were used to ignore accidental hyperparameters and models with no generalization ability caused by extraordinary partitions, so as to enhance the generalization ability of the model. The data set was divided into ten subsets, one of which was taken as the validation set, and the rest was taken as the training set. During the process, the hyperparameters were kept stable to measure their advantages and disadvantages, and the ultimately obtained hyperparameters were used to train the entire data model in all data.

While keeping the hyperparameters consistent, the average training loss and the average validation loss of the 10 models were taken to measure the hyperparameters. After the models were built, the first 30 features were input into the models to generate air quality predictions and further analyze air quality changes. The performance evaluation index of the regression model can measure the deviation degree of the forecast results of continuous values from the real data. We compare the predicted values of ten verification processes with the actual calculated concentrations. Methods including Mean Absolute Error (MAE), the coefficient of determination (R2), and the Root Mean Square Error (RMSE) were used to represent the difference between the label and the predicted value. The smaller these values, the better the performance of the regression model, and the predicted results are closer to the ground truth level.

We train this model with 30 features and $PM_{2.5}$ labels, and Grid Search CV and Randomized Search CV are commonly used for hyperparameter optimization. Grid Search CV is a straightforward procedure that tries each set of hyperparameters one by one and selects the best one. This approach consumes too much time resources, so Randomized Search CV (RSCV) is chosen as an alternative in this study, and the introduction of a random factor can improve the efficiency of the optimization search in some cases, saving computational time by using only a fixed number of parameter settings to find the locally optimal solution.

2.8. Estimating AQ Mappings with Gridded Networks. The ML-LUR proposed in this study replaces the model part with XGBoost and combines satellite AOD inversion, meteorological parameters, and land use type parameters to estimate $PM_{2.5}$ concentrations near the ground. Due to the relative uncertainty of remote sensing satellite products, aerosol thickness inversion results were combined with the global solar photometer network to render high-precision AOD measurements with an uncertainty of less than 0.02. The atmospheric chemical transport model (GEOS-Chem) simulated AOD is also used as a significant fraction of the AOD source. Satellite observations comprise 89% of the global population-weighted AOD data from December 1, 2019, to December 1, 2021. Ultimately, the mapping of spatiotemporal AQ was achieved by using information on meteorological parameter characteristics (temperature, relative humidity, and wind speed), land use type characteristics, digital elevation, and latitude and longitude as model feature inputs.

3. Results

3.1. The Annual Seasonal Variation of $PM_{2.5}$. The variation trend of fine particulate air pollutants in China from December 2019 to November 2021 is shown in Table 2, in which the overall concentration change of $PM_{2.5}$ and aerosol thickness, the average seasonal change, and the annual average change trend are presented by year and quarter. The overall average $PM_{2.5}$ concentration and AOD values are 33.77 (SD = 31.15) $\mu g/m^3$ and 0.08 (SD = 0.20). In the seasonal breakdown, the average $PM_{2.5}$ concentrations are in the following order from highest to lowest: winter, spring, autumn, and summer. In winter, the average $PM_{2.5}$ concentration can reach 53.50 (SD = 40.43) $\mu g/m^3$, while in summer, the average $PM_{2.5}$ concentration is only 19.04 (SD = 12.36) $\mu g/m^3$, while the AOD values are larger in winter and spring (0.10) and the smallest in summer (0.04). This inconsistency between the magnitude and season of the data for $PM_{2.5}$ concentration and AOD may be due to other factors (e.g., weather factors and human factors). The last two rows of Table 1 show the annual average $PM_{2.5}$ concentrations and AOD values for 2020 and 2021, respectively. Compared to 2020, the average $PM_{2.5}$ concentration in 2021 decreases from 34.26 (SD = 31.43) $\mu g/m^3$ to 30.95 (SD = 28.86) $\mu g/m^3$, which is a decrease in the pollution level.

In this study, because of the large spatial latitude and longitude span in China, it is relatively normal that there will be a less close association between $PM_{2.5}$ concentrations and AOD values aggregated between different regions of the country. For example, in northwest China, where the temperature difference between day and night is large, and air particles are generally an important indicator of atmospheric stratification; we observed that the vertical distribution of aerosols adapts to the thermal change of the boundary layer, resulting in the low height of the atmospheric mixed layer, and the stratification of the atmosphere occurs in a very short time in a day. Therefore, the research shows that the relationship between the characteristics of the short-term mixing layer and pollutant concentration is not so close in the northwest plateau area. On the contrary, in southern China, the temperature difference between day and night is generally considerable, and the occurrence of atmospheric stratification takes a comparatively long time. Similarly, it can be determined that the characteristics of the short-term mixed layer in this region are more closely related to the study of air pollutant particles, and the vertical distribution of aerosol is more closely related to $PM_{2.5}$ concentration.

We also found low correlations for simple linear regressions for most of the coastal areas throughout the study period. The main reason for this phenomenon may be that the coastal areas are affected by numerous types of near-surface wind, which are mostly uneven and deflected, and under these conditions, pollutants are transported between land and ocean. In addition, due to the particularity of the weather in coastal areas, which are greatly influenced by ocean currents and sea and land winds, as well as a variety of complex terrain comparable to hills, pollutant diffusion conditions are frequently quite different from those in inland areas, resulting in the phenomenon that emission sources, AOD, and pollutant concentration do not constitute a simple linear proportion. In addition, the influence of ocean climate on clouds will also make aerosol observation problematic and lead to a certain degree of error. Meanwhile, most of the monitoring sites in the coastal areas of the country show correlations below the overall average correlation. The seasonal transport pattern of air quality may lead to relatively low correlations due to the more active air mixing in different seasons in coastal areas, as well as the strong influence of the local sea breeze.

Table 2: Average/annual/quarterly $PM_{2.5}$ measurement and AOD inversion from remote sensing data from December 2019 to November 2021 in China.

	$PM_{2.5}$ concentration ($\mu g/m^3$)	AOD
Overall	33.77 (31.15)	0.08 (0.20)
Season		
Winter	53.50 (40.43)	0.10 (0.21)
Spring	33.08 (30.73)	0.10 (0.24)
Summer	19.04 (12.36)	0.04 (0.16)
Fall	29.70 (23.37)	0.07 (0.18)
Year		
2020	34.26 (31.43)	0.08 (0.20)
2021	30.95 (28.86)	0.08 (0.20)

Table 3 gives the overall, seasonal, and annual model performance for China during the study period. The R2 values of the overall training model were 0.88, MAE was 7.56, RMSE was 15.51, and SMAPE was 20.62. The R2 values of the training model were above 0.80 for all four quarters and above 0.86 for all quarters except summer, where the training model had the highest R2 in autumn (0.90), followed by winter (0.88), spring (0.86), and summer (0.80). The fit differences among the four seasons are small, indicating that the fitted model can explain the $PM_{2.5}$ concentration changes better.

The RMSE values of the model were the largest in winter (14.70), followed by spring (9.91), fall (7.40), and summer (4.83); the magnitudes of MAE values were 8.29 (winter), 5.16 (spring), 3.41 (summer), and 4.69 (fall); and the magnitudes of SMAPE values were 17.27 (winter), 17.07 (spring), 20.40 (summer), and 19.58 (fall), respectively. Both MAE and RMSE values reflect the error before the true and fitted values, so the magnitudes are consistent. Considering that the model uses a large number of sample data nationwide and also the variability of data from different regions is large, the maximum error is relatively small and the simulation results are reliable by the true error reflected by MAE and the amplified error reflected by RMSE.

The R2 value of the training model in 2020 is 0.89, and the MAE, RMSE, and SMAPE are 5.63, 10.95, and 18.58, respectively; the R2 value of the training model in 2021 is 0.90, and the MAE, RMSE, and SMAPE are 4.93, 8.58, and 18.28, respectively. The R2 values of the two training models are high, indicating that the models are well fitted; the three sets of error results are relatively close, indicating that the

Table 3: Description of the results and evaluation for the ML-LUR model.

	N	R2	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	SMAPE (%)
Overall	925357	0.88	7.56	15.51	20.62
(A) Season					
Winter	227604	0.88	8.29	14.70	17.27
Spring	232623	0.86	5.16	9.91	17.07
Summer	228551	0.80	3.41	4.83	20.40
Fall	236582	0.90	4.69	7.40	19.58
(B) Year					
2020	470071	0.89	5.63	10.95	18.58
2021	415159	0.90	4.93	8.58	18.28

predicted values are closer to the true values before, and the sample data of the two time periods are more consistent.

3.2. Spatial Mappings of $PM_{2.5}$ Concentrations. Figure 2 shows the spatial distribution of the average estimated $PM_{2.5}$ concentrations in China from December 2019 to November 2021. Since the AOD is retrieved under cloud-free conditions, the spatial pattern of mean estimated $PM_{2.5}$ is more likely to represent $PM_{2.5}$ levels on cloud-free days, which are more common during the warm season. This spatial distribution map shows that the average $PM_{2.5}$ concentrations are higher in the northwest and north China plain regions, with an average of about 40 $\mu g/m^3$, especially in the southern region of Xinjiang, which is basically above 50 $\mu g/m^3$; the average $PM_{2.5}$ concentrations in the southeast coastal region are more consistently distributed, mostly between 20 $\mu g/m^3$ and 30 $\mu g/m^3$; the environmental quality in the Tibetan region is better, with average $PM_{2.5}$ concentrations generally below 10 $\mu g/m^3$.

The results demonstrated that the spatial contrast of PM2.5 was influenced by the broad spatial distribution coverage rate, land utilization ratio and coverage rate, and the terrain difference in different areas. The northern part of China is comparatively flat and has a lot of plain terrain, so pollutants in the air are more readily dispersed, and the regional transport efficiency is relatively high. At the same time, anthropogenic factors such as agricultural pollution, industrial pollution, and automobile exhaust combined with the weak purification function of the ecosystem make air pollution in Northwest China serious; the main pollution in Northwest China is sand and dust, which is affected by extreme weather and wind whenever it cools down, and dusty weather can lead to a high $PM_{2.5}$ index, while weak cold air activity, low precipitation, and long duration of still windy weather during the heating period are not conducive to pollutant diffusion; the continuous accumulation of pollutants aggravates the pollution level; the Tibetan region has high forest coverage and is located in the plateau, the land is wide and sparse, and few pollutants are emitted.

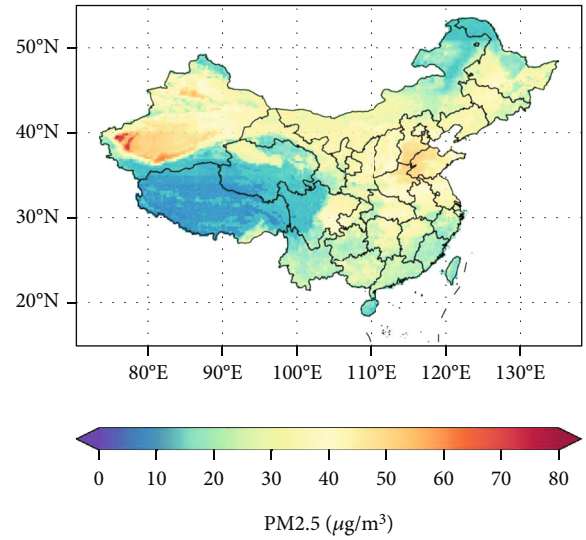

Figure 2: Spatial map of estimated $PM_{2.5}$ concentrations in China for the period of December 2019-November 2021.

4. Conclusion

In this study, we used the AOD data provided by MODIS and the predictors such as pollution factors and land use factors extracted by the land use regression (LUR) model to build models for different seasons based on the machine learning algorithm XGBoost to achieve the prediction of the spatial distribution of $PM_{2.5}$ concentrations near the whole China. The results show that the training model fits well, with the R2 values of the fitting coefficients of 0.86 (spring), 0.80 (summer), 0.90 (autumn), and 0.88 (winter), and the LUR model based on the XGBoost algorithm can effectively reduce the spatial heterogeneity of geographic variables and explain more than 80% of the variation of $PM_{2.5}$ concentration values. By comparing the simulated and real values, the results show that the accuracy of the validation data set is relatively high, with an average error of no more than 3%, indicating that the prediction model can effectively estimate the spatial distribution of near-ground $PM_{2.5}$ concentrations across the country and explain the characteristics of $PM_{2.5}$ concentration distribution.

Prediction results to some extent show that the present situation of air pollution situation is still significant, through operative to estimate the concentration distributions of nationally; on the one hand, it can provide the shortages of pollutant monitoring method of improving scientific

guidance and improve the layout of the base station monitoring and numerous auxiliary means such as satellite monitoring to achieve the purpose of precise monitoring. On the other hand, analyzing the formation factors of pollutants through the difference of regional pollutant concentration is conducive to reducing pollutant emission from the root.

Data Availability

The air quality data are collected from the China Environment Monitoring Center.

Conflicts of Interest

All authors disclosed no relevant relationships.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62172061; National Key Research and Development Project under Grant Nos. 2020YFB1711800 and 2020YFB1707900; Science and Technology Project of Sichuan Province under Grant Nos. 2021-YFG0152, 2021YFG0025, 2020YFG0479, 2020YFG0322, 2020GFW035, and 2020GFW033; and R&D Project of Chengdu City under Grant No. 2019-YF05-01790-GX.

References

[1] T. Suwa, J. C. Hogg, K. B. Quinlan, A. Ohgami, R. Vincent, and S. F. van Eeden, "Particulate air pollution induces progression of atherosclerosis," *Journal of the American College of Cardiology*, vol. 39, no. 6, pp. 935–942, 2002.

[2] A. Harper, P. N. Baker, Y. Xia et al., "Development of spatiotemporal land use regression models for $PM_{2.5}$ and NO_2 in Chongqing, China, and exposure assessment for the CLIMB study," *Pollution Research*, vol. 12, no. 7, article 101096, 2021.

[3] M. H. Forouzanfar, A. Afshin, L. T. Alexander et al., "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1659–1724, 2016.

[4] A. Nemmar, P. H. M. Hoet, B. Vanquickenborne et al., "Passage of inhaled particles into the blood circulation in humans," *Circulation*, vol. 105, no. 4, pp. 411–414, 2002.

[5] M.-A. Kioumourtzoglou, J. D. Schwartz, M. G. Weisskopf et al., "Long-term $PM_{2.5}$ exposure and neurological hospital admissions in the northeastern United States," *Environmental Health Perspectives*, vol. 124, no. 1, pp. 23–29, 2016.

[6] R. M. Hoff and S. A. Christopher, "Remote sensing of particulate pollution from space: have we reached the promised land?," *Journal of the Air & Waste Management Association*, vol. 59, no. 6, pp. 645–675, 2009.

[7] H. Zhang, R. M. Hoff, and J. A. Engel-Cox, "The relation between moderate resolution imaging spectroradiometer (MODIS) aerosol optical depth and $PM_{2.5}$ over the United States: a geographical comparison by U.S. Environmental Protection Agency regions," *Journal of the Air & Waste Management Association*, vol. 59, no. 11, pp. 1358–1369, 2009.

[8] J. Engel-Cox, C. H. Holloman, B. W. Coutant, and R. M. Hoff, "Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality," *Atmospheric Environment*, vol. 38, no. 16, pp. 2495–2509, 2004.

[9] H. J. Lee, Y. Liu, B. A. Coull, J. Schwartz, and P. Koutrakis, "A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations," *Atmospheric Chemistry and Physics*, vol. 11, no. 15, pp. 7991–8002, 2011.

[10] Y. Liu, J. A. Sarnat, V. Kilaru, D. J. Jacob, and P. Koutrakis, "Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing," *Environmental Science & Technology*, vol. 39, no. 9, pp. 3269–3278, 2005.

[11] A. van Donkelaar, R. V. Martin, M. Brauer et al., "Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application," *Environmental Health Perspectives*, vol. 118, no. 6, pp. 847–855, 2010.

[12] D. Vienneau, K. de Hoogh, M. J. Bechle et al., "Western European land use regression incorporating satellite- and ground-based measurements of NO_2 and PM_{10}," *Environmental Science & Technology*, vol. 47, no. 23, pp. 13555–13564, 2013.

[13] Q. Di, H. Amini, L. Shi et al., "An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution," *Environment International*, vol. 130, article 104909, 2019.

[14] A. Shtein, I. Kloog, J. Schwartz et al., "Estimating daily $PM_{2.5}$ and PM_{10} over Italy using an ensemble model," *Environmental Science & Technology*, vol. 54, no. 1, pp. 120–128, 2019.

[15] A. M. Sayer, N. C. Hsu, C. Bettenhausen, and M.-J. Jeong, "Validation and uncertainty estimates for MODIS collection 6 "deep blue" aerosol data," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 14, pp. 7864–7872, 2013.

[16] M. Stafoggia, T. Bellander, S. Bucci et al., "Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013-2015, using a spatiotemporal land-use random- forest model," *Environment International*, vol. 124, pp. 170–179, 2019.

[17] Y. Liu, C. J. Paciorek, and P. Koutrakis, "Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information," *Environmental Health Perspectives*, vol. 117, no. 6, pp. 886–892, 2009.

[18] A. W. Strawa, R. B. Chatfield, M. Legg, B. Scarnato, and R. Esswein, "Improving retrievals of regional fine particulate matter concentrations from moderate resolution imaging spectroradiometer (MODIS) and ozone monitoring instrument (OMI) multisatellite observations," *Journal of the Air & Waste Management Association*, vol. 63, no. 12, pp. 1434–1446, 2013.

[19] Z. Ma, X. Hu, L. Huang, J. Bi, and Y. Liu, "Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing," *Environmental Science & Technology*, vol. 48, no. 13, pp. 7436–7444, 2014.

[20] W. Song, H. Jia, J. Huang, and Y. Zhang, "A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the Pearl River Delta region in China," *Remote Sensing of Environment*, vol. 154, pp. 1–7, 2014.

[21] A. Chudnovsky, H. J. Lee, A. Kostinski, T. Kotlov, and P. Koutrakis, "Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the geostationary operational environmental satellite (GOES)," *Journal of the Air & Waste Management Association*, vol. 62, no. 9, pp. 1022–1031, 2012.

[22] I. Kloog, P. Koutrakis, B. A. Coull, H. J. Lee, and J. Schwartz, "Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements," *Atmospheric Environment*, vol. 45, no. 35, pp. 6267–6275, 2011.

[23] G. Hoek, R. Beelen, K. de Hoogh et al., "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric Environment*, vol. 42, no. 33, pp. 7561–7578, 2008.

[24] M. T. Young, M. J. Bechle, P. D. Sampson et al., "Satellite-based NO_2 and model validation in a national prediction model based on universal kriging and land-use regression," *Environmental Science & Technology*, vol. 50, no. 7, pp. 3686–3694, 2016.

[25] E. Dons, M. van Poppel, B. Kochan, G. Wets, and L. Int Panis, "Modeling temporal and spatial variability of traffic-related air pollution: hourly land use regression models for black carbon," *Atmospheric Environment*, vol. 74, pp. 237–246, 2013.

[26] A. Saraswat, J. S. Apte, M. Kandlikar, M. Brauer, S. B. Henderson, and J. D. Marshall, "Spatiotemporal land use regression models of fine, ultrafine, and black carbon particulate matter in New Delhi, India," *Environmental Science & Technology*, vol. 47, no. 22, pp. 12903–12911, 2013.

[27] B. Hellack, D. Sugiri, R. P. F. Schins et al., "Land use regression modeling of oxidative potential of fine particles, NO_2, $PM_{2.5}$ mass and association to type two diabetes mellitus," *Atmospheric Environment*, vol. 171, pp. 181–190, 2017.

[28] R. M. J. Beelen, G. Hoek, D. Vienneau et al., "Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe - The ESCAPE project," *Atmospheric Environment*, vol. 72, pp. 10–23, 2013.

[29] Y. Tian, X. Yao, and L. Chen, "Analysis of spatial and seasonal distributions of air pollutants by incorporating urban morphological characteristics," *Computers, Environment and Urban Systems*, vol. 75, pp. 35–48, 2019.

[30] J. Song and M. Stettler, "A novel multi-pollutant space-time learning network for air pollution inference," *Science of The Total Environment*, vol. 811, article 152254, 2022.

[31] J. Xu, W. Yang, B. Han et al., "An advanced spatio-temporal model for particulate matter and gaseous pollutants in Beijing, China," *Atmospheric Environment*, vol. 211, pp. 120–127, 2019.

[32] A. I. Lyapustin, Y. Wang, I. Laszlo et al., "Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm," *Journal of Geophysical Research*, vol. 116, no. D3, 2011.

[33] L. Chen, Z. Bai, S. Kong et al., "A land use regression for predicting NO_2 and PM_{10} concentrations in different seasons in Tianjin region, China," *Journal of Environmental Sciences*, vol. 22, no. 9, pp. 1364–1373, 2010.

[34] L. Kong and G. Tian, "Assessment of the spatio-temporal pattern of $PM_{2.5}$ and its driving factors using a land use regression model in Beijing, China," *Environmental Monitoring and Assessment*, vol. 192, no. 2, p. 95, 2020.

[35] S. A. A. Shah, W. Aziz, M. Almaraashi, M. S. A. Nadeem, N. Habib, and S.-O. Shim, "A hybrid model for forecasting of particulate matter concentrations based on multiscale characterization and machine learning techniques," *Mathematical Biosciences and Engineering: MBE*, vol. 18, no. 3, pp. 1992–2009, 2021.

[36] B. Zhang, H. Zhang, G. Zhao, and J. Lian, "Constructing a $PM_{2.5}$ concentration prediction model by combining auto-encoder with Bi-LSTM neural networks," *Environmental Modelling & Software*, vol. 124, article 104600, 2020.

[37] N. Chen, M. Yang, W. du, and M. Huang, "$PM_{2.5}$ estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution MAIAC AOD in Hubei, China," *ISPRS International Journal of Geo-Information*, vol. 10, no. 1, p. 31, 2021.

[38] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. Zhang, "Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2285–2297, 2018.

[39] T. Zheng, M. H. Bergin, R. Sutaria, S. N. Tripathi, R. Caldow, and D. E. Carlson, "Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in Delhi," *Atmospheric Measurement Techniques*, vol. 12, no. 9, pp. 5161–5181, 2019.

[40] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, San Francisco, California, USA, August 2016.