Hindawi
Journal of Sensors
Volume 2023, Article ID 9836421, 1 page
https://doi.org/10.1155/2023/9836421

Retraction

Retracted: Optimization of Land Use Regression Modelling of $PM_{2.5}$ Spatial Variations in Different Seasons across China

Journal of Sensors

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Chai, J. Song, L. Zhang, B. Guo, and Y. Xu, "Optimization of Land Use Regression Modelling of $PM_{2.5}$ Spatial Variations in Different Seasons across China," *Journal of Sensors*, vol. 2022, Article ID 3659254, 9 pages, 2022.

Hindawi
Journal of Sensors
Volume 2022, Article ID 3659254, 9 pages
https://doi.org/10.1155/2022/3659254

Research Article

Optimization of Land Use Regression Modelling of $PM_{2.5}$ Spatial Variations in Different Seasons across China

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

[1]*College of Computer Science, Sichuan University, China*
[2]*Chengdu Hankang Information Industry co., LTD, China*
[3]*Department of Geography, Faculty of Social Science, 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 10 August 2022; Revised 31 August 2022; Accepted 16 September 2022; Published 28 September 2022

Academic Editor: Yuan Li

Fine particulate matter ($PM_{2.5}$), one of the main components of haze, is of wide concern for its potential negative health effects. In order to further improve ambient air quality, it is essential to conclude the spatial variability of pollutants by investigating air pollution exposure. We divide China into two parts, north and south, and use a Land Use Regression (LUR) model to extract data including meteorological data, land use factors, and AOD retrievals, and use the machine learning algorithm to optimize the model to achieve predictions of the spatial distribution of near-surface $PM_{2.5}$ mass concentrations in southern and northern China. We evaluated the seasonal consistency of the models in southern and northern China, and in northern China, we found a better fit with better seasonal consistency for the heating season and annual average model, while in southern China, we did not find a more fitted seasonal phase. The study illustrates that it is feasible to simulate the spatial distribution of $PM_{2.5}$ mass concentration in large-scale areas based on the LUR model, and the seasonal consistency of the LUR model has been done to some extent.

1. Introduction

With the rapid development of China's economy, the rapid urbanization has not only improved material wealth and living standards but also caused environmental impacts, especially the increasingly serious air pollution situation, and the phenomenon of haze pollution in large areas and serious standards have appeared in every place. Relevant monitoring data show that China has now become one of the most serious $PM_{2.5}$ pollution areas in the world [1]. In early 2013, hazy weather hit half of China, covering an area of more than 1.3 million square kilometers at the worst time, with persistent heavy pollution in many cities [2]. Air pollution not only seriously harms people's health, causing a large number of respiratory diseases and even death, but also causes flight delays, factory closures, and other social problems, which affect social and economic development.

Fine particulate matter can enter the human body through the respiratory system and body fluid system, which is more harmful to the human body and attracts more attention. Moreover, compared with other large suspended particulate matter, $PM_{2.5}$ can be suspended in the air for a longer time, which is easier to enter the human system and cause cardiovascular and lung diseases [3]. In EU countries, $PM_{2.5}$ leads to an average reduction in life expectancy of 8.6 months [4]. More than two million premature deaths worldwide are now linked to particulate pollutants. $PM_{2.5}$ was first identified as a carcinogen by the World Health Organization in 2013 [5].

Researching the spatial variation of pollutants can essentially help us study the transport of urban air pollutants and the health effects of exposure to urban pollutants, such as spatial interpolation, atmospheric chemical transport models, atmospheric diffusion models, and other atmospheric pollutant estimation methods [6]. Land use regression (LUR) model has become an important method to predict the long-term spatial distribution of pollutant concentration and is widely used in urban air pollution prediction. Compared with other model prediction methods, LUR can explain the spatial distribution of air pollutant

concentration from the perspective of influence mechanism [7–11]. The research and application of land use regression modelling optimization of $PM_{2.5}$ spatial variation in different seasons in China provide methodological experience and theoretical basis for population exposure, epidemiological research, and health risk assessment [12–15]. The LUR model is a general model to simulate the spatial differentiation of air pollutant concentrations at the urban scale. It is an empirical model that can be used to establish large areas on the urban scale. It usually adopts a modelling model combining various statistical methods to establish the statistical relationship between the covariates of air pollutant observation data and geostatistical information. Based on land use information, the atmospheric pollutant concentration in unmonitored areas is anticipated by regression [16, 17]. By establishing land use information, traffic conditions, and population distribution around the site as predictors, the LUR model can be applied to estimate the pollutant concentration in any region at multiple spatial scales when the change of the existing pollutant concentration is fully explained [18–20]. At a small spatial scale, such as urban areas, the concept of using the same set of predictor variables to forecast the concentration of pollutants in all areas of the research area is evidently flawed because of the differences in major pollution sources in distinct areas. Compared with the method of air diffusion model, the LUR model based on remote sensing data provides background concentration supplement for urban spatial scale, compensates for the prediction defect caused by incomplete predictor variables to a certain extent, and can better characterize the spatial variation trend of pollutant concentration [21–23].

The traditional LUR model fits the relationship between the predictors and the concentration of air pollutants, usually using the statistical method of linear regression. In order to describe the possible uncertain nonlinear correlation, in this study, we implement machine learning-based LUR model to achieve spatiotemporal AQ inference to construct a high-resolution grid-based AQ mappings by exploiting meteorological conditions, land use variables, AOD inversion results, and pollutant concentrations at air quality monitoring stations. By analyzing high-resolution grid-based AQ mappings, we explore the space-time pollutant concentration variations and show the characteristics of the results based on observed data sets across China between 2019 and 2021. We employ a novel feature engineering approach to construct a robust model in consideration of space-time effect and heterogeneity, which combines the advantages of land use variables, AOD inversion results based on remote sensing satellite, and meteorological variables. The state-of-the-art machine learning algorithm, XGBoost, will be incorporated as the surrogate model in our study to train the model based on over 30 million observations from meteorological stations, 1593 samples from air quality monitoring stations, and more than 380,000 times aerosol optical depth inversion based on remote sensing satellite. The results will benefit our knowledge about the air pollution situation and policy-making for air pollution control.

The air pollutant data collected from fixed monitoring stations were simulated and studied. The land use regression model was used to generate a fine-scale spatial variation grid of fine particulate matter concentration distribution on a nationwide scale [24–26]. Air quality monitoring results show that hourly pollutant concentration varies in different locations and time periods. Although the pollutant concentration data required for research can be retrieved by means of multisource monitoring, which is dynamic and hourly, the variables used in land use generally only need static average data [27]. The LUR model is used to study the simulation of small-scale spatial pollutant concentration changes by using average sampled pollutant concentration data to describe the pollution characteristics of an area and the individual exposure levels of the population in that area [28–31]. According to some air quality monitoring studies, LUR model produces different results in different periods in the same area, and season is the main influencing factor affecting the concentration of atmospheric pollutant particles [32–34].

LUR model mainly reflects the information of pollution sources and their diffusion conditions. In addition to land use information and season, the independent variables considered by the model also comprehensively consider traffic, industrial emissions, meteorological conditions, topography, population distribution, and other factors, which can fully reflect the spatial differentiation of small-scale pollutant concentration. In other words, LUR models are normally time restricted, and their valid time is usually only for the period of time that the model is operating. It remains to be researched whether the model can be used as a consistent model for the prediction results of annual average and daily average concentration in the following time period in the study area.

In this study, we analyzed the spatial distribution of air pollution in China and divided the whole geographical region into two parts, north and south, with Qinling-Huaihe River as the dividing line. The northern region has the characteristic of centralized heating in winter, while the southern region does not have this characteristic, as shown in Figure 1. Based on the characteristic of centralized heating in northern China, we will combine XGBoost to fit the annual average LUR model. For the south, we will also combine XGBoost to fit the spring, summer, fall, and winter seasons and the annual mean LUR model to compare the seasonal consistency of the LUR model.

2. Materials and Measurements

2.1. Study Area. The Qinling-Huaihe Line is commonly used as the geographical boundary between the north and the south, and its geographical conditions regularly restrict the convection of gases between the north and the south, resulting in the difference of climatic conditions between the north and the south. Divided by this boundary, the northern part of the country has freezing winters, so there is usually regional and seasonal central heating, so there are obvious differences in climatic conditions and geographical environment between the north and south of the dividing line, ranging roughly from 31° to 35° north latitude, as shown in Figure 1. North of Qinling-Huaihe Line is the geographical

Figure 1: Heating situation map distribution of China.

division of northern China. Rivers and lakes are cold and dry in winter, lakes freeze, annual precipitation is low, precipitation is short and mainly concentrated in summer, river water volume is small, and the water level changes greatly. On the contrary, the situation in the south of Qinling-Huaihe Line is just the opposite. There, the rivers generally do not freeze in winter, and the climate is mild and little rain, so the river water volume is large and the water level changes little. The north will have central heating in winter, usually between November and March, while the south will have no central heating.

2.2. Air Quality Data. We collected historical data on air pollution concentrations from December 2019 to December 2021. To study the LUR modelling of $PM_{2.5}$ distribution changes, we conducted a comparative study on seasonality in northern China, designated the period from November to March of the next year as the heating season, and the rest of the period as the nonheating season. Because central heating varies by region and environment temperature, the time of cold wave will not be the same every year due to the climate difference. Therefore, central heating areas need to wait for a specific heating time to use the heating system.

2.3. Ground-Level AQ Measurements. We used real-time ground monitoring air quality information from the National Environmental Monitoring Center in China and selected official monitoring site data from the government to ensure the quality of the data. Monitoring indicators include the city's daily air quality index, the city's hourly air quality index, and the average hourly concentration of $PM_{2.5}$ released at points. The concentration of $PM_{2.5}$ pollutants was recorded, and the hourly data recorded at the monitoring sites were used to estimate the daily average. A total of more than 1600 static monitoring stations in China were selected, and daily monitoring data of more than 15 records of each monitoring station were selected to ensure the stability of the computed data. Monitoring indicators include the city's daily air quality index, the city's hourly air quality index, and the average hourly concentration of $PM_{2.5}$ released at points. The hourly concentration of particulate matter at the monitoring point is the arithmetic mean or measured value of the concentration measured at the point within one hour.

Because the published results are usually updated every hour and the data transmission takes a certain amount of time, the published data will be delayed. In addition, some monitoring sites have instrument calibration or routine maintenance activities in part of the time, so some sites will have data loss for a period of time. We removed the extreme (abnormally high or minus values) AQ data samples and fill the missing values by sliding window method.

2.4. Satellite AOD Retrievals. Data were obtained using 1 km resolution terrestrial aerosol optical depth 2 data from NASA's remote sensing data product MCD19A2. Generally, the inversion effect of aerosol is greatly affected by the weather, and there will be serious data loss when there are clouds in the observation area of satellite orbit. Because MCD19A2 combines two satellite data, the data loss rate is lower than those of other atmospheric satellite data products, and its algorithm is advanced with high spatial and temporal resolution. The MAIAC algorithm extracts spectral regression coefficients from the time series of satellite images and realizes aerosol inversion and bidirectional surface reflectance based on multiangle. The algorithm uses time series observation data. The results of Dark Target and Deep Blue algorithms are better than those of Dark Target and Deep Blue algorithms [35]. MCD19A2 data were provided by the NASA Center, and a collection of daily MCD19A2 from Dec. 2019 to Nov. 2021 was used in this study.

To achieve global coverage within China's geographical range, 22 orbit data in MODIS were used. The MCD19A2 Version 6 data is in HDF format, and the data is converted to TIF format by the MRT (MODIS Reprojection Tool) provided by NASA, and the AOD bands in the data are selected to obtain the mean values and perform nationwide splicing. In this study, the AOD value extracted from the $0.01^\circ \times 0.01^\circ$ range scale at each $0.25^\circ \times 0.25^\circ$ grid center is considered as the representative estimate of each grid.

2.5. Gridded Meteorology Data. The assimilated meteorological data are from the Global Tropospheric Analyses and Forecast Grids dataset spanning from December 2019 to December 2021, which is provided by National Centers for Environmental Prediction (NCEP). The grid is $0.25^\circ \times 0.25^\circ$ meteorological parameter information, such as wind speed and direction and temperature. The above

operational data are aggregated to daily means for modelling in this study.

2.6. Feature Engineering Approach. Static and dynamic features are selected for AQ modelling, as listed in Table 1, including land use covers, meteorological parameters, AOD retrievals, location attributes (longitude and latitude), and time attributes. We extracted five variables from the meteorological parameters, including wind condition, pressure, temperature, and RH. Here, we develop a novel feature engineering approach by extracting the higher correlated feature variables which enhance the model capability to achieve more robust and reliable inference. We select one third of the total data and order them according to their feature importance as the training set.

3. Design of Frameworks

3.1. Development of ML-LUR Model. XGBoost is an optimized parallel distributed gradient enhancement library designed to be efficient, flexible, and portable. Based on the gradient propulsion framework, it implements tree propulsion in parallel. It is a highly scalable system with sparse sensing, which can solve various data science problems quickly and accurately.

XGBoost is based on a gradient lifting mechanism. The basic idea is to screen out the sample features as the classifier model, minimize the objective function through residual learning, and repeatedly generate multiple simple models to form a new complex model. The core of the new model is to control the complexity of the model while establishing the gradient direction of the corresponding loss function and correcting the residual. In addition, XGBoost is more efficient than neural networks, which is very convenient for frequent parameter optimization in experiments.

XGBoost is an enhanced version of GBDT. Compared with GBDT, its algorithm is mainly improved in regularization promotion and parallel distribution. Adding regular terms into the objective function can effectively reduce the structural risk of the model and prevent overfitting. In addition, XGBoost supports parallelism in feature granularity, so multithreading can be used to calculate the optimal segmentation point of each feature during node splitting to reduce computer memory consumption. These improvements have greatly improved the training speed of XGBoost and expanded the application range of its algorithm.

This is a supervision model based on regression tree, and its objective function is

$$\mathrm{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{1}$$

The formula contains two parts: error function and regularization term. The error function uses cross entropy, and the regularization term is superimposed by the regularization term of K trees, which is helpful to smooth the final learning weight and can effectively avoid overfitting. The regularization term of the KTH tree is as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2. \tag{2}$$

In the formula, γ and λ are model parameters, w_j is the weight of the j-th node in the tree, and w_j uses L2 norm to better avoid overfitting.

3.2. Estimating AQ Mappings with Gridded Networks. AERONET is an aerosol remote sensing observation network developed by NASA, the network now covers major regions of the world with more than 500 sites. CIMEL automatic solar photometer (SPAM) was used as the basic observation instrument. AERONET plays an important role in studying the radiative transfer mode and verification of global aerosols. AERONET is complementary to satellite remote sensing, and the optical thickness measured by AERONET is usually used as ground-truth to test the accuracy of aerosol optical thickness retrieved by remote sensing.

We develop the ML-LUR using XGBoost model and produce ground-based estimates of surface AQ concentrations exploiting a combination of satellite AOD retrievals, meteorological parameters, and land use configurations. We combine satellite data products with AERONET for high-precision aerosol measurements, and as the northern region is sparsely covered by satellites, we use GEOS-Chem simulation aerosol data as an additional data source supplement. We then incorporate the gridded meteorological variables (e.g., temperature and RH) and land use configurations together as features to recover space-time AQ mappings.

4. Results

4.1. Descriptive Statistics for $PM_{2.5}$ Concentration Data. Since the start of environmental control efforts, including the closure of heavy polluting enterprises and the installation of pollutant filtering devices, $PM_{2.5}$ pollution has been greatly improved. But surveys from 2019 to 2021 show that more than 90 percent of the population is still exposed to areas where the average annual $PM_{2.5}$ concentration exceeds the national standard of 15 $\mu g/m^3$, while the proportion of people exposed to areas where the average annual $PM_{2.5}$ concentration exceeds the national standard of 35 $\mu g/m^3$ is still more than 60 percent. From a global perspective, about 30% of the population exposed to the average $PM_{2.5}$ concentrations exceeded primary standard, but in North America and Europe and other developed countries and regions of the $PM_{2.5}$ exposure ratio keep normal level. We find the $PM_{2.5}$ problems in China still keep serious and need to contribute more in the air pollution controlling work. What is more, it is valuable to conduct more statistic work for the north and south regions in China.

4.1.1. Statistics in Northern China. In northern China, there will be centralized government heating in winter, and its common heating season is from November to March. We collected historical data on air pollution concentrations from

Table 1: Supporting features used in this study.

Data category		Data type
Static	Geographic and land use information	Digital elevation (m)
		Area of cultivated land (km^2)
		Area of wood land (km^2)
		Area of waters (km^2)
		Area of ocean (km^2)
Dynamic	Meteorology information	Wind speed (m/s)
		Pressure (kPa)
		Temperature (°C)
		Specific humidity (kg/m^3)
		Relative humidity (%)

December 2019 to December 2021 and saw that the calculated results of the average $PM_{2.5}$ concentration in China are 38.83 $\mu g/m^3$ and 21.11 $\mu g/m^3$ in the heating season and nonheating season in the northern region, and the overall annual mean concentration is 28.45 $\mu g/m^3$ in the northern.

4.1.2. Statistics in Southern China. In southern China, we fit the LUR model to the southern region of China with four seasons and build the corresponding models for four seasons, respectively. The average annual concentration in each season was 23.54$\mu g/m^3$ in spring, 13.15 $\mu g/m^3$ in summer, 19.78 $\mu g/m^3$ in autumn, and 33.88 $\mu g/m^3$ in winter. The average $PM_{2.5}$ concentration in the southern region are 30.77 $\mu g/m^3$ and 16.73 $\mu g/m^3$ in the heating season and nonheating season, and the overall annual mean concentration is 22.55 $\mu g/m^3$ in the southern.

4.2. Exploration of ML-Based LUR Modelling

4.2.1. Space-Time Modelling in Northern China. We calculated the model and found that the LUR model is mainly affected by these following factors: (1) month, (2) latitude, (3) longitude, (4) specific humidity, (5) AOD, and (6) relative humidity during the heating season in northern China. The R^2 of the LUR model is 0.90, the RMSE is 14.31 $\mu g/m^3$, the SMAPE is 16.88%, and the MAE is 8.59 $\mu g/m^3$.

In the nonheating season in northern China, the LUR model is mainly affected by these following factors: (1) month, (2) longitude, (3) latitude, (4) AOD, (5) temperature, and (6) specific humidity. The R^2 of the LUR model is 0.7952, the RMSE is 10.36 $\mu g/m^3$, the SMAPE is 19.50%, and MAE is 4.72 $\mu g/m^3$.

The annual average $PM_{2.5}$ LUR model for northern China is mainly affected by these following factors: (1) month, (2) latitude, (3) longitude, (4) specific humidity, (5) AOD and (6) pressure. The R^2 of the model is 0.85, RMSE is 12.75 μg/m3, SMAPE is 19.67%, and MAE is 6.66 $\mu g/m^3$.After that, we fit the heating season LUR model, the nonheating season LUR model, and the annual average $PM_{2.5}$ LUR model for comparison. There is a strong consistency between the main influencing variables of the annual average model and the time of the heating season model. It can be seen from the R^2 index results that the model index results in the heating season are better than those in the nonheating season, in which the R^2 result of the heating season is 0.91, and the $R^2$of the nonheating season model is 0.795. The results indicate that the annual average spatial pattern of $PM_{2.5}$ is mainly influenced by the pollution in the heating season.

4.2.2. Space-Time Modelling in Southern China. In spring, the LUR model is mainly affected by these following factors: (1) month, (2) longitude, (3) latitude, (4) wood land, (5) relative humidity, and (6) AOD retrievals. The R^2 of the model is 0.85, RMSE is 12.75 $\mu g/m^3$, SMAPE is 19.67%, and MAE was 6.66 $\mu g/m^3$.

In summer, the LUR model is mainly affected by these following: factors (1) latitude, (2) longitude, (3) month, (4) wood land, (5) temperature, and (6) water covers. The R^2 of the model is 0.75, the RMSE is 3.99 $\mu g/m^3$, the SMAPE is 19.12%, and the MAE is 2.80 $\mu g/m^3$.

In autumn, the LUR model is mainly affected by these following factors: (1) month, (2) latitude, (3) longitude, (4) specific humidity, (5) temperature, and (6) AOD. The R^2 of the model is 0.8769, RMSE is 6.21 $\mu g/m^3$, SMAPE is 18.84%, and MAE was 4.20 $\mu g/m^3$.

In winter, the LUR model is mainly affected by these following factors: (1) latitude, (2) longitude, (3) month, (4) ELEVATION, (5) AOD, and (6) waters. The R^2 of the model is 0.89, RMSE is 8.74 $\mu g/m^3$, SMAPE is 15.95%, and MAE is 5.99 $\mu g/m^3$.

The annual average LUR model is mainly influenced by these following factors: (1) month, (2) latitude, (3) longitude, (4) specific humidity, (5) AOD, and (6) ELEVATION. The $R^2$of the model is 0.89, RMSE is 6.58 $\mu g/m^3$, SMAPE is 18.37%, and MAE is 4.19 $\mu g/m^3$.

We train and fit different LUR models based on spring, summer, fall, winter and annual average level for model comparision. In summer and winter, we find that dimension is the most influential factor with the highest weight, which may be caused by the huge difference between summer and winter climate, where the weather is hot and rainy in summer and cold and dry in winter, and the difference of dimension determines the difference of weather climate. In spring and autumn, month is the first determinant, and even in

autumn its weight factor is more than 50%, and the second ranking is longitude. Based on the model analysis we can see that there is no season in which the grid map fitted by the model matches the annual average $PM_{2.5}$ grid map. The results show that the annual average model of LUR has poor temporal consistency with the four seasonal models of spring, summer, autumn, and winter.

4.3. Machine Learning-Based LUR Mapping. Although natural gas has gradually been used as winter heating energy in recent years, coal still dominates, especially in northern cities. Coal produces more $PM_{2.5}$ precursors (suspended particles that are formed through chemical reactions) than natural gas. In order to study the influence of northern heating season on $PM_{2.5}$ and eliminate the influence of meteorological conditions on pollutant concentration, we integrated meteorological data as the benchmark variable to establish pollutant regression model.

4.3.1. Space-Time Modelling in Northern China. We established a LUR model for simulating the whole country to intuitively evaluate the spatial distribution characteristics of $PM_{2.5}$ and analyzed the spatial characteristics of regional concentration. As shown in Figure 2(a), the figure shows the concentration prediction spatial distribution in the nonheating season, and Figure 2(b) figure shows the concentration prediction spatial distribution in the heating season. The results demonstrate that the spatial distribution characteristics of $PM_{2.5}$ in northern China are more evident, and the overall pollutant concentration in the heating season is much higher than that in the nonheating season, so the heating supply has a significant impact on air pollution.

4.3.2. Space-Time Modelling in Southern China. In order to analyze the spatial distribution of $PM_{2.5}$ in southern China more intuitively, the spatial distribution of $PM_{2.5}$ concentration values in southern China was simulated for four quarters shown as Figure 3. We can tell from the results in the figure that $PM_{2.5}$ in southern China shows obvious spatial distribution characteristics, and the average spatial distribution of $PM_{2.5}$ in each season is different. The highest concentration of $PM_{2.5}$ is in winter, and the concentration of $PM_{2.5}$ in the remaining seasons from high to low is in spring, autumn, and summer.

5. Discussion

$PM_{2.5}$ concentrations are known to be higher in northern China during the heating season than those in other seasons, and one of the main reasons for this is the presence of a temperature inversion, which is comparable to creating a "cover" over the region. The increase in near-surface pollutant emission is not conducive to the horizontal regional transportation of pollutants, resulting in the formation of haze phenomenon in the region [36, 37]. In addition to the influence of meteorological conditions, the elevated emission of pollutants in the region is also an important factor.

During the heating season, municipalities focus on burning coal for heating, and currently the whole northern region relies on coal combustion for heating, and in many, places poor quality coal is used. In addition, the incomplete combustion of fuel oil is also one of the important reasons for the increase of pollutant emission during winter heating. Many studies have shown that coal combustion is an important source of $PM_{2.5}$. Studies have shown that about 30% of $PM_{2.5}$ comes from direct emissions from coal combustion, motor vehicles, dust, etc. (primary particulate matter), and 70% is converted to particulate matter (secondary particulate matter). Therefore, China has recently taken measures to retrofit and upgrade its heating to reduce pollution emissions at the source, such as eliminating inefficient boilers and using clean energy [38].

In the southern region, even though there is no centralized government provision of heating in the southern region in winter, the $PM_{2.5}$ concentration in southern China is significantly higher in the winter season than those in other seasons, which may be due to the influence of winter climate. In general, the temperature decreases with the increase of altitude, the lower air is hotter and the upper air is colder, the cold air is heavy and sinks, and the hot air is light and will rise, forming convection. However, in winter, the ground temperature decreases, resulting in the atmospheric structure above the ground will appear the temperature increases with the height of the "inverse temperature" phenomenon. Once this cold inversion layer is formed, the air cannot be converted up and down, and it is difficult for pollutants to spread. At the same time, frequent rainfall and blowing weather make the atmospheric haze in summer be cleaned to a certain extent. However, due to the dry weather in winter, there is rarely rainfall, and the rainfall is low, so the reduction of the cleaning ability of the natural environment is also one of the reasons for the high concentration of $PM_{2.5}$ in winter.

We can see that in southern China, $PM_{2.5}$ concentrations vary significantly seasonally. In the southern region, the first influence parameter of the annual average $PM_{2.5}$ model is month, and even in the autumn LUR model, the weighting factor of month reaches 0.7921; this shows that the concentration of $PM_{2.5}$ is sensitive to the season. In summer and winter, we find that dimension is the influence factor with the largest weight, which may be caused by the huge climate difference between summer and winter, and the difference of dimension determines the difference of weather climate.

According to the $PM_{2.5}$ grid map, we can see the urban centers with the highest pollution levels, such as the Beijing-Tianjin-Hebei region, while the border areas of the cities tend to have relatively low pollutant concentrations. The research and analysis of the spatial scale distribution of pollutants at the city scale are often based on the choice of urban residents and road construction pattern planning, which is related to the rapid urbanization development in China. If the central area of a city develops rapidly and its living conditions are better than those in the marginal areas, there will be a higher distribution of residents, and people's daily and business activities tend to be concentrated in the urban center. However, people's requirements for the living environment are gradually increasing. Under the higher air environment requirements, the government's urban planning definitely tends to move industrial activities to the

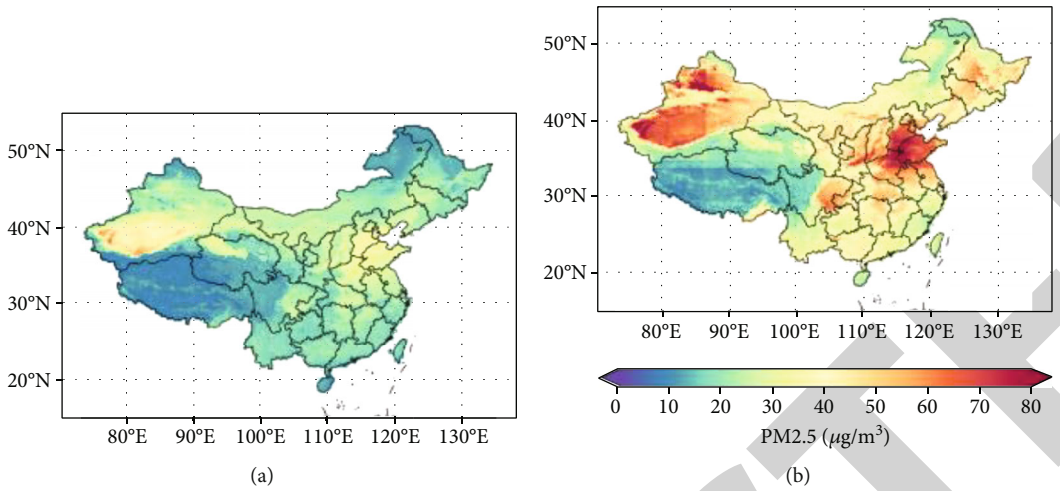

Figure 2: Grid diagram of LUR model for nonheating season from April to October (a) and heating season from November to March (b).

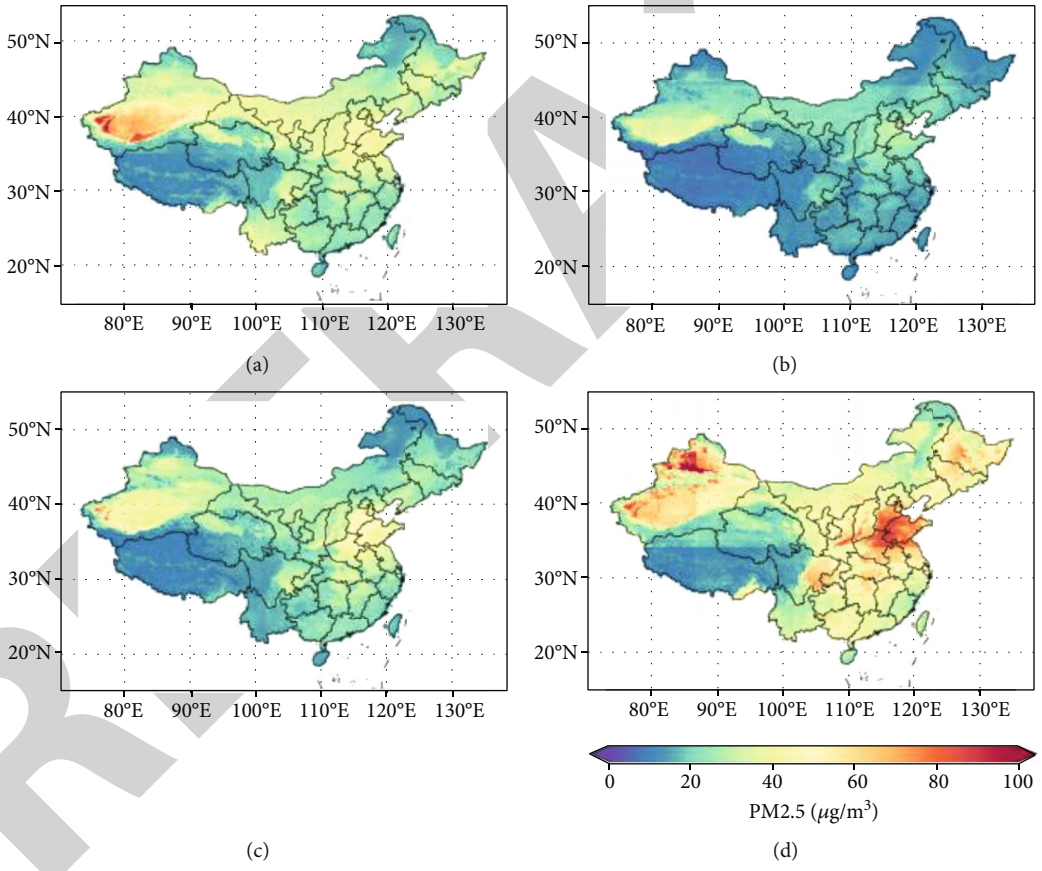

Figure 3: The results of seasonal gridded concentration prediction model are spring (a), summer (b), autumn (c), and winter (d).

urban fringe areas, so the concentration level in the city center will theoretically alleviate to a certain extent in the future [39, 40].

6. Conclusion and Prospect

We studied the spatial distribution of $PM_{2.5}$ in China and divided China into two parts, north and south, for the different characteristics of south and north China, we used the machine learning algorithm XGBoost and land use regression model combined with meteorological data, land use factors and AOD data to predict the spatial distribution of $PM_{2.5}$ concentrations in different seasons. The most important predictor of the spatial variation of $PM_{2.5}$ concentration is month. In the northern region of China, the model fitting result of heating season ($R^2 = 0.8992$) was better than that of

nonheating season ($R^2 = 0.7952$). The results indicate that the LUR model for the heating season is in good temporal agreement with the annual average model, and the annual average spatial pattern of $PM_{2.5}$ is mainly influenced by the pollution in the heating season. In southern China, the R^2 of the LUR model for the spring, summer, autumn, and winter seasons were 0.8489, 0.7468, 0.8879, and 0.8927, respectively, with the highest average $PM_{2.5}$ in winter. We did not find a better agreement between the LUR model and the annual average LUR model in which season in southern China.

This paper studies and discusses the spatial and temporal distribution characteristics of air pollutant $PM_{2.5}$ due to geographical differences and seasonal alternation in northern and southern China. Through the forecast of natural conditions and the study of the impact of human social behavior on environmental pollution, it is profitable to provide scientific guidance for reducing air pollution. And it has long-term implications for economic-driven analysis and the study of diseases related to human health. In the future, richer prediction models can be constructed by considering more diverse influencing factors, such as the inherent association with air pollution and quantification of economic losses under the current research conditions of novel topics such as COVID19 or new energy mix.

Data Availability

The air quality data are collected from China Environmental 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 Program of China under Grant No.2020YFB1711800 and 2020YFB1707900; the Science and Technology Project of Sichuan Province under Grant No.2021-YFG0152, 2021YFG0025, 2020YFG0479, 2020YFG0322, 2020GFW035, 2020GFW033, and the Research and Development Project of Chengdu City under Grant No.2019-YF05-01790-GX.

References

[1] L. J. Han, W. Q. Zhou, and W. F. Li, "City as a major source area of fine particulate ($PM_{2.5}$) in China," *Environmental Pollution*, vol. 206, pp. 183–187, 2015.

[2] R. Burnett, H. Chen, M. Szyszkowicz et al., "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 38, pp. 9592–9597, 2018.

[3] X. Du, Q. Kong, W. Ge, S. Zhang, and L. Fu, "Characterization of personal exposure concentration of fine particles for adults and children exposed to high ambient concentrations in Beijing, China," *Journal of Environmental Sciences-China*, vol. 22, no. 11, pp. 1757–1764, 2010.

[4] P. Prendes, J. M. Andrade, P. López-Mahía, and D. Prada, "Source apportionment of inorganic ions in airborne urban particles from Coruña City (N.W. of Spain) using positive matrix factorization," *Talanta*, vol. 49, no. 1, pp. 165–178, 1999.

[5] D. Loomis, Y. Grosse, and B. Laubysecretan, "The International Agency for Research on Cancer monograph working group IARC," *The Carcinogenicity of Outdoor Air Pollution*, vol. 1, 2013.

[6] D. Briggs, A. Aaheim, C. Dore, G. Hoek, M. Petrakis, and G. Shaddick, *Air Pollution Modelling for Support to Policy on Health and Environmental Risks in Europe*, APMoSPHERE Final Report, 2005.

[7] M. D. Adams and P. S. Kanaroglou, "Mapping real-time air pollution health risk for environmental management: combining mobile and stationary air pollution monitoring with neural network models," *Journal of Environmental Management*, vol. 168, pp. 133–141, 2016.

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

[9] M. Eeftens, R. Beelen, K. D. Hoogh et al., "Development of land use regression models for PM2.5, PM2.5absorbance, PM10and PMcoarsein 20 European study areas; results of the ESCAPE project," *Environmental Science & Technology*, vol. 46, no. 20, pp. 11195–11205, 2012.

[10] C. C. Ho, C. C. Chan, C. W. Cho, H. I. Lin, J. H. Lee, and C. F. Wu, "Land use regression modeling with vertical distribution measurements for fine particulate matter and elements in an urban area," *Atmospheric Environment*, vol. 104, pp. 256–263, 2015.

[11] M. Lee, M. Brauer, P. Wong et al., "Land use regression modelling of air pollution in high density high rise cities: a case study in Hong Kong," *Science of the Total Environment*, vol. 592, pp. 306–315, 2017.

[12] T. Shi, N. Dirienzo, W. J. Requia, M. Hatzopoulou, and M. D. Adams, "Neighbourhood scale nitrogen dioxide land use regression modelling with regression kriging in an urban transportation corridor," *Atmospheric Environment*, vol. 223, article 117218, 2020.

[13] J. Lee, C. Wu, G. Hoek et al., "LUR models for particulate matters in the Taipei metropolis with high densities of roads and strong activities of industry, commerce and construction," *The Science of the Total Environment*, vol. 514, pp. 178–184, 2015.

[14] I. Aguilera, J. Sunger, R. Ferandez-Patier et al., "Estimation of outdoor NOx, NO2, and BTEX exposure in a cohort of pregnant women using land use regression modeling," *Environmental Science & Technology*, vol. 42, no. 3, pp. 815–821, 2008.

[15] E. Dons, M. V. Poppel, B. Kochan, G. Wets, and L. I. Panis, "Modeling temporal and spatial variability of traffic-related air pollution: hourly land use regression models for black carbon," *Atmospheric Environmental*, vol. 74, no. 2, pp. 237–246, 2013.

[16] D. J. Briggs, S. Collins, P. Elliott et al., "Mapping urban air pollution using GIS: a regression-based approach," *International Journal of Geographical Information Science*, vol. 11, no. 7, pp. 699–718, 1997.

[17] 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.

[18] P. Hystad, E. Setton, A. Cervantes et al., "Creating national air pollution models for population exposure assessment in Canada," *Environmental Health Perspectives*, vol. 119, no. 8, pp. 1123–1129, 2011.

[19] C. Liu, B. H. Henderson, D. F. Wang, X. Y. Yang, and Z. R. Peng, "A land use regression application into assessing spatial variation of intra- urban fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations in City of Shanghai, China," *Science of the Total Environment*, vol. 565, pp. 607–615, 2016.

[20] H. Tularam, L F. Ramsay, S. Muttoo, B. Brunekreef, and R N. Naidoo, "A hybrid air pollution / land use regression model for predicting air pollution concentrations in durban, south africa," *Environmental Pollution*, vol. 7, p. 116513, 2021.

[21] M. B. Dijkema, U. Gehring, R. T. van Strien et al., "A comparison of different approaches to estimate small-scale spatial variation in outdoor NO2concentrations," *Environmental Health Perspectives*, vol. 119, no. 5, pp. 670–675, 2011.

[22] J. D. Marshall, E. Nethery, and M. Brauer, "Within-urban variability in ambient air pollution: comparison of estimation methods," *Atmospheric Environment*, vol. 42, no. 6, pp. 1359–1369, 2008.

[23] M. Wang, U. Gehring, G. Hoek et al., "Air pollution and lung function in Dutch children: a comparison of exposure estimates and associations based on land use regression and dispersion exposure modeling approaches," *Environmental Health Perspectives*, vol. 123, no. 8, pp. 847–851, 2015.

[24] L. Chen, S. Gao, H. Zhang et al., "Spatiotemporal modeling of $PM_{2.5}$ concentrations at the national scale combining land use regression and Bayesian maximum entropy in China," *Environment International*, vol. 116, pp. 300–307, 2018.

[25] J. Kerckhoffs, M. Wang, K. Meliefste et al., "A national fine spatial scale land-use regression model for ozone," *Environmental Research*, vol. 140, pp. 440–448, 2015.

[26] Z. Y. Zhang, J. B. Wang, J. E. Hart et al., "National scale spatiotemporal land-use regression model for $PM_{2.5}$, PM_{10} and NO_2 concentration in China," *Atmospheric Environment*, vol. 192, pp. 48–54, 2018.

[27] E. Dons, M V. Poppel, L I. Panis, S D. Prins, and C. Matheeussen, "Land use regression models as a tool for short, medium and long term exposure to traffic related air pollution," *Science of The Total Environment*, vol. 476-77C, pp. 378–386, 2014.

[28] J. Song, K. Han, and M. Stettler, "Deep-MAPS: machine learning based mobile air pollution sensing," *IEEE Internet of Things Journal*, vol. 8, pp. 7649–7660, 2021.

[29] X U. Xia, Q. Gao, C. Peng, X. Cui, Y. Liu, and J. Li, "Integrating global socio-economic influences into a regional land use change model for china," *Frontiers of Earth Science*, vol. 8, no. 1, pp. 81–92, 2014.

[30] E M. Noth, S K. Hammond, G S. Biging, F. Lurmann, and I B. Tager, "Mixed modeling for land use regression with traffic-related pollutants," *Epidemiology*, vol. 19, no. 6, p. S327, 2008.

[31] L. Weissert, J. Salmond, G. Miskell, M. Alavi-Shoshtari, and D. Williams, "Development of a microscale land use regression model for predicting NO_2 concentrations at a heavy trafficked suburban area in Auckland, NZ," *Science of the Total Environment*, vol. 619-620, pp. 112–119, 2018.

[32] 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," *Science of the Total Environment*, vol. 22, no. 9, pp. 1364–1373, 2010.

[33] Y. Ghassoun, M. Ruths, M. O. Lowner, and S. Weber, "Intra-urban variation of ultrafine particles as evaluated by process related land use and pollutant driven regression modelling," *Science of the Total Environment*, vol. 536, pp. 150–160, 2015.

[34] S. Mukerjee, R. D. Willis, J. T. Walker et al., "Seasonal effects in land use regression models for nitrogen dioxide, coarse particulate matter, and gaseous ammonia in Cleveland, Ohio," *Atmospheric Pollution Research*, vol. 3, no. 3, pp. 352–361, 2012.

[35] M. Tao, J. Wang, R. Li et al., "Performance of MODIS high-resolution MAIAC aerosol algorithm in China: characterization and limitation," *Atmospheric Environment*, vol. 213, pp. 159–169, 2019.

[36] Z. Q. Li, J. P. Guo, A. J. Ding et al., "Aerosol and boundary-layer interactions and impact on air quality," *National Science Review*, vol. 4, no. 6, pp. 810–833, 2017.

[37] Y. R. Yang, X. G. Liu, Y. Qu et al., "Formation mechanism of continuous extreme haze episodes in the megacity Beijing, China, in January 2013," *Atmospheric Research*, vol. 155, pp. 192–203, 2015.

[38] Y. Zong, Q. Zhang, C. Hong, and K. He, "Assessment of the benefits of emission reductions from coal-fired source emission control measures in Beijing," *Research of Environmental Sciences*, vol. 30, pp. 1645–1652, 2017.

[39] Y. Chen, N. Schleicher, Y. Chen, F. Chai, and S. Norra, "The influence of governmental mitigation measures on contamination characteristics of $PM_{2.5}$ in Beijing," *Science of the Total Environment*, vol. 490, pp. 647–658, 2014.

[40] Y. Zhou, Y. Wu, L. Yang et al., "The impact of transportation control measures on emission reductions during the 2008 Olympic Games in Beijing, China," *Atmospheric Environment*, vol. 44, no. 3, pp. 285–293, 2010.