RESEARCH ON PM2.5 MASS CONCENTRATION RETRIEVAL METHOD BASED ON HIMAWARI-8 IN BEIJING

Fulin Luo 1, Juanli Jing 1,2, *, Anna Wang 1, Lisi Liang 1

1 College of Geomatics and Geoinformation, Guilin University of Technology
2 Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin 541004, China

KEY WORDS: Himawari-8; AOD; PM2.5; MLR; GWR

ABSTRACT:

This paper was based on Japan's new generation of geostationary satellite Himawari-8 2016 Aerosol Optical Depth (AOD) data and near-ground monitoring station PM2.5 mass concentration data, boundary layer height (BLH), relative humidity (RH), normalized vegetation index (NDVI) data to establish a multivariate linear regression model (MLR) and a geographically weighted regression model (GWR) in Beijing. This provided data and scientific basis for the treatment of air pollution. The results show that: (1) The fitting determination coefficient R² of the MLR was 0.5244, indicating that there was a significant correlation between PM2.5 and AOD. After GWR model introduced BLH, RH and NDVI in turn, R² increased from 0.3945 to 0.5403, indicating that the introduction of relevant influencing factors can improve the accuracy of the model, that was, PM2.5 was affected by BLH, RH and NDVI. (2) The regression coefficients of the MLR and GWR of the BLH, RH and NDVI were statistically analyzed. The regression coefficients of the two models were close to each other, but the standard deviation of the GWR regression coefficients was larger than the MLR, indicating that the local information of the GWR model was more abundant. It reflected the difference characteristics of the regression coefficients of each parameter.

1. INTRODUCTION

Aerosol is characterized by solid and liquid particles suspended in the atmosphere. Aerosol optical Depth (AOD) is an important parameter of aerosols and is defined as the integral of the extinction coefficient of the medium in the vertical direction. Fine particles with a dynamic diameter of no more than 2.5μm are called PM2.5, which is the main pollutant causing heavy air pollution (Cao, 2014). At present, there are two ways to monitor the mass concentration of PM2.5, one is ground-based monitoring and the other is satellite remote sensing. Compared with satellite remote sensing, ground monitoring has higher accuracy, but the spatial coverage of monitoring stations is limited and has certain limitations (Hu et al., 2013). Compared with the traditional ground monitoring methods, satellite remote sensing has the advantage of regional scale, which can reflect the macroscopic distribution trend of pollutants in large space range and different scales (Hu, et al., 2015; Tao et al., 2012). Research scholars at home and abroad have conducted research on satellite remote sensing monitoring of PM2.5 mass concentration: Chen et al. (2014) used the MODIS AOD data and quoted the relative humidity (RH) and boundary layer height (BLH) to construct the relative humidity of the partition. The results show that the zonal relative humidity method has certain research value for studying the correlation between AOD and PM2.5. Zhang et al. (2016) established a semi-physical geographic weighted regression model (GWR) to estimate PM10 by vertical correction and humidity correction of AOD and PM2.5. The results showed that the corrected GWR can improve the correlation between PM10 and AOD. The coefficient of determination R² increased from 0.08 to 0.43. Li et al. (2018) combined principal component analysis with GWR to construct a geo-weighted model based on principal component analysis (PCA-GWR). The results showed that PCA-GWR was more accurate than GWR. It performed principal component analysis on multiple...
related variables, which can play a role in dimensionality reduc-
tion. Wang et al. (2017) used the AOD product of the
Himawari-8 satellite and used AERONET AOD as the true
value, which confirmed the high precision of the Himawari-8
AOD data and can effectively characterize the change of
aerosol optical thickness. It can be applied to the AOD-PM2.5
correlation analysis of Beijing-Tianjin-Hebei. The mixed effect
model was used to estimate the PM2.5 concentration per hour,
which proved that Himawari-8 AOD product had certain value
in the study of pm2.5 mass concentration. Previous studies have
suggested that the correlation between AOD and PM2.5 varies
with the spatial environment. Different geographical regions
had different aerosol types and GWR was proposed to better
limit the spatial heterogeneity of large-scale
regression (Fotheringham et al. 1996; Li et al. 2016; Na et
al. 2010; Zhang et al. 2015). Traditional research on AOD and
PM2.5 basically used MODIS AOD products with a spatial
resolution of 10 km and 3km, while Himawari-8 AOD has
higher temporal resolution and spatial resolution than MODIS
AOD. Therefore, this paper estimated the PM2.5 mass
centration in Beijing based on the Japanese meteorological
satellite Himawari-8 2016 AOD data, and introduced the
boundary layer height (BLH), relative humidity (RH), and
normalized vegetation index (NDVI). Multivariate linear
regression model and geographic weighted regression model
(GWR) were used to analyze the correlation between AOD and
PM2.5 and compare the applicability of the two models to air
pollution problems in Beijing.

1. STUDY AREA AND DATASETS

2.1 Study Area

Beijing is located at 115.7°E-117.4°E and 39.4°N-41.6°N. It is
located in the north of the north China plain, backed by
Yanshan, adjacent to Tianjin and Hebei Province, and has
jurisdiction over 16 districts including Dongcheng, Xicheng,
Chaoyang, Haidian, Fengtai and Shijingshan. The climate of
Beijing is typical north temperate subhumid continental
monsoon climate, with high temperatures and rains in summer
and cold and dry winters. Beijing has a large and dense
population, large demand for resources, and large pollutant
emissions. The topography of Beijing's three sides is not
conducive to the spread of pollutants, making foggy weather
often occur in Beijing, and the problem of air pollution is very
serious.

2.2 Research Data

2.2.1 Himawari-8 Meteorological Satellite: The new
generation of Japanese geostationary meteorological satellite
Himawari-8 carries the most advanced optical sensor AHI. It
has 16 observation bands with a spatial resolution of 0.5 or 1
km in the visible and 2 km in near-infrared bands. These
advantages, combined with reduced revisit time (approximately
10 minutes in total and approximately 2.5 minutes in the sector),
provide new levels of competence for identifying and tracking
rapidly changing weather phenomena and for deriving
quantitative products (Bessho et al. 2016). Himawari-8 can
provide AOD products with spatial resolutions of 5km, 1km,
2km and time resolution: 10min, 1h. It covers about one-third of
the Earth's range (ie West Pacific, East Asia and Southeast Asia,
Oceania). This paper used AOD data with a temporal resolution
of 1h and a spatial resolution of 5km.

2.2.2 PM2.5 Mass Concentration: The PM2.5 mass
concentration data used in this paper was from the PM2.5
hourly average concentration data of ground monitoring from
January 2016 to December 2016 in Beijing air quality historical
data (http://beijingair.sinaapp.com/). The 35 PM2.5 ground
monitoring sites in Beijing are shown in Figure 1.

2.2.3 Meteorological Parameters: Meteorological data,
including Relative Humidity (RH) and Boundary layer height

Figure 1. Distribution map of Beijing PM2.5 monitoring stations
The meteorological data used in this paper was from the reanalysis data set of the European Centre For Medium-Range Weather Forecasts (ECMWF) (ie ERA-Interim). ERA Temporary, one of the ECMWF reanalyzed data sets has provided global climate reanalysis since 1979.

2.2.4 Surface Parameters: This article obtained the MODIS 16-day 1km resolution "MOD13A2" and "MYD13A2" normalized difference vegetation index (NDVI) from the MODIS official website (https://ladsweb.modaps.eosdis.nasa.gov/). It was processed to get 8 days of data. NDVI greater than 0.4 usually indicates vegetated areas, while smaller values usually indicate soil-dominated soil surfaces (Liu et al., 2014).

2.2.5 Data Matching: In this paper, the above data was processed and matched by MATLAB programming for modeling, verification and analysis. In space, AOD, meteorological parameters and surface parameters were averaged by non-zero pixels with a radius of 5 km centered on the PM2.5 site. In terms of time, the data were averaged from 9-16 in Beijing time from January 1, 2016 to December 31, 2016 as the daily average.

2. RESEARCH METHODS

3.1 Multiple Linear Regression Model

In the real world, the change of the dependent variable is often affected by several important factors. At this time, it is necessary to use two or more influencing factors as the independent variables to explain the change of the dependent variable. This is called multiple regression or multiple regression. When there is a linear relationship between multiple independent variables and dependent variables, the regression analysis carried out is pluralistic regression. Let y be the dependent variable, \( x_1, x_2, ..., x_k \) are independent variables, and when there is a linear relationship between the independent variable and the dependent variable, then the multiple linear regression model is:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k
\]  
(1)

Where \( \beta_0 \) is the constant term, \( \beta_1, \beta_2, ..., \beta_k \) are regression coefficients.

3.2 Geographically Weighted Regression Model

Geographically weighted regression is a new spatial analysis method proposed in recent years. It is used to detect the non-stationarity of spatial relationships by embedding spatial structures in linear regression models. The geographically weighted regression model used in this study is as follows:

\[
y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \epsilon_i \quad i = 1, 2, ..., n
\]  
(2)

Where \( (u_i, v_i) \) are the coordinates of the ith sample point (such as latitude and longitude), \( \beta_k(u_i, v_i) \) is the kth regression parameter at the ith sample point, which is a function of the geographic position.

\( \epsilon_i \) is an independent and identically distributed random error, usually assumed to obey \( N(0, \sigma^2) \).

The spatial weight matrix is the core of the geographically weighted regression model. The spatial position of the data is embedded in the regression parameters, and the point-by-point parameter estimation is performed by the local weighted least squares method. Weight is the distance function between the geospatial position of the regression point and the geospatial position of other observation points. Common spatial weight function methods are: distance threshold method, distance inverse ratio method, Gauss function method, among which the universal application is Gaussian function method. Its function form is as follows (Xuan et al., 2014):

\[
W_{ij} = \exp\left( -\left( \frac{d_{ij}}{b} \right)^2 \right)
\]  
(3)

Where \( d_{ij} \) is the distance between sample points i and j.

b is a non-negative attenuation parameter that describes the functional relationship between weight and distance.

The larger the bandwidth, the slower the weight decays with distance, and the smaller the bandwidth, the faster the weight decays with distance. The AIC information criterion (ie, Akaike information criterion) can be used to measure the goodness of statistical model fitting. The formula of AIC is as follows:
where \( tr(S) \) = the trace of the matrix \( S \) of GWR, which is a function of the bandwidth \( b \).

\[ AIC = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n\left[\frac{n + \text{tr}(S)}{n - 2 - \text{tr}(S)}\right] \]  

(4)

Where \( \text{tr}(S) \) = the trace of the matrix \( S \) of GWR, which is a function of the bandwidth \( b \).

\( n \) = the size of the sample

\( \hat{\sigma} \) = the maximum likelihood estimation of random error term variance, i.e., for the same sample data, the AIC value is made. The bandwidth corresponding to the smallest geographically weighted regression function is the optimal bandwidth.

3. RESULTS

3.1 Variable Statistical Histogram

Figure 2. mainly illustrated the histogram statistics of the variables used in the model. Overall, in addition to the annual average NDVI and RH showed a bimodal distribution, the overall trend of the remaining variables showed a lognormal distribution. In 2016, the ground PM2.5 concentration ranged from 3μg/m\(^3\) to 513μg/m\(^3\), and its annual average value was 67.9249μg/m\(^3\), exceeding the secondary standard of ambient air quality (35μg/m\(^3\)). The annual average of AOD was 0.3668 and the standard deviation was 0.2854. The BLH was mainly distributed between 400m and 1500m, and the maximum and minimum values were 3510.1m and 126.1392m, respectively, and the average value was 1177.5m. RH was mainly distributed between 22% and 60%, with a maximum of 102.5811% and a minimum of 9.6727%. NVDI was mainly distributed between 0.1 and 0.7, with a maximum of 0.8419 and a minimum of -0.162.

3.2 Results of Multiple Linear Regression Model

Table 1. showed the output of the MLR model, in which PM2.5 was \( y \) (dependent variable), and AOD, BLH, RH, and NDVI were \( x_1, x_2, x_3, \) and \( x_4 \) (independent variables), respectively. It can be seen from Table 1. that \( R^2=0.5244, F=43.65, p=0.0000, \) and \( p<0.05, \) the regression model \( y=90.643+142.836x_1-0.028x_2-0.769x_3-27.678x_4 \) could be obtained.
3.3 Results of Geographic Weighted Regression Model

Table 2. was a scatter plot of PM2.5 predicted values and near-surface PM2.5 observations obtained by adding different parameters (including AOD, BLH, RH, NDVI) to the GWR model. After adding the parameters in the GWR model, the coefficient $R^2$ was determined to increase from 0.3945 to 0.5403, indicating that meteorological factors affected the spatial distribution of pollutants. After introducing more suitable parameters, the accuracy of the model was improved, that was, the PM2.5 mass concentration accuracy predicted by the model was improved.

<table>
<thead>
<tr>
<th>ID</th>
<th>N</th>
<th>parameter</th>
<th>AIC</th>
<th>$R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>6180</td>
<td>AOD</td>
<td>65,038.60</td>
<td>0.3945</td>
<td>29.3</td>
</tr>
<tr>
<td>b</td>
<td>3905</td>
<td>AOD, RH</td>
<td>40,898.85</td>
<td>0.4333</td>
<td>30.7</td>
</tr>
<tr>
<td>c</td>
<td>6180</td>
<td>AOD, NDVI</td>
<td>64,472.68</td>
<td>0.4496</td>
<td>29.38</td>
</tr>
<tr>
<td>d</td>
<td>3703</td>
<td>AOD, BLH</td>
<td>38,733.87</td>
<td>0.4878</td>
<td>31.66</td>
</tr>
<tr>
<td>e</td>
<td>3905</td>
<td>AOD, RH, NDVI</td>
<td>40,692.76</td>
<td>0.4645</td>
<td>30.37</td>
</tr>
<tr>
<td>f</td>
<td>3703</td>
<td>AOD, BLH, NDVI</td>
<td>38,627.53</td>
<td>0.5045</td>
<td>31.41</td>
</tr>
<tr>
<td>g</td>
<td>3405</td>
<td>AOD, BLH, RH</td>
<td>35,336.75</td>
<td>0.5356</td>
<td>31.7</td>
</tr>
<tr>
<td>h</td>
<td>3405</td>
<td>AOD, BLH, RH, NDVI</td>
<td>35,314.17</td>
<td>0.5403</td>
<td>31.72</td>
</tr>
</tbody>
</table>

Table 2. Comparison of fitting results of different parameters in the model

3.4 Comparative Analysis of Model Results

Figure 3. showed the results of the MLR and GWR h models. The slope of the MLR model was 0.52, $R^2$ was 0.52, and RMSE was 37.6. The slope of the GWR model was 0.54, $R^2$ was 0.54, and RMSE was 31.7. Overall, the GWR model fit slightly better than the MLR, but it was not obvious. The reason might be that the scope of the Beijing research area was small, and the monitoring stations within the research scope were relatively close and the number was small, so the effective monitoring data volume was insufficient and the data difference was weak in the data processing process.
The average values of the regression coefficients of the statistical MLR model and the GWR model were shown in Table 3. From the table, the regression coefficients of the two models were generally close. The PM2.5 mass concentration was generally proportional to AOD and inversely proportional to RH, BLH, and NDVI, indicating that the GWR model maintains overall consistency with the MLR model. It was worth noting that from the standard deviation of each regression coefficient, the standard deviation of the regression coefficient of the GWR model was larger than that of the MLR model. This showed that the local information of the GWR model was richer, and it can better reflect the difference characteristics of the regression coefficients of each parameter.

<table>
<thead>
<tr>
<th>Regression parameter</th>
<th>MLR</th>
<th>GWR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>regression coefficient</td>
<td>standard deviation</td>
</tr>
<tr>
<td>Beta_AOD</td>
<td>142.8360</td>
<td>2.7770</td>
</tr>
<tr>
<td>Beta_RH</td>
<td>-0.7963</td>
<td>0.0512</td>
</tr>
<tr>
<td>Beta_BLH</td>
<td>-0.0280</td>
<td>0.0013</td>
</tr>
<tr>
<td>Beta_NDVI</td>
<td>-27.6781</td>
<td>5.9645</td>
</tr>
</tbody>
</table>

Table 3. Statistics of geographically weighted and multiple linear regression parameters

In order to further analyze the spatial difference of the regression coefficients, the spatial distribution of the regression results in winter was calculated according to the GWR results (see the Figure 4). It could be seen from Figure 4. that the regression parameters all showed obvious regional distribution characteristics. The overall performance of Beta_AOD was the highest in Cheng 8 and Shunyi districts, and the lowest in Fangshan district. Beta_RH indicated the coefficient of moisture absorption, and Miyun, Pinggu, Shunyi and Huairou were relatively largest, and Daxing District was relatively smallest. Beta_BLH showed the regional difference of PM2.5 vertical distribution structure to some extent. Beta_NDVI reflected the background influence of plant canopy, and its spatial distribution was relatively high in the city's eight districts and relatively low in other regions.
4. CONCLUSION

Based on the meteorological data and ground observation data of the Himawari-8 satellite AOD products from November 2016 to December 2016, the correlation analysis between AOD and PM2.5 was carried out by MLR and GWR. The results showed that:

1. The MLR model determination coefficient $R^2$ to be 0.4316 and the root mean square error was 47.07, indicating that there was a significant correlation between PM2.5 and AOD. With the introduction of meteorological parameters into the GWR model, the determination coefficient $R^2$ of the linear correlation analysis between the predicted GWR model values and ground observations increased from 0.3945 to 0.5403, indicating that the accuracy of the model was affected by meteorological factors and other parameters. From the perspective of goodness of fit, the GWR model was superior to the MLR. And after introducing relevant factors, the accuracy of the model can be improved, and future research will continue to improve the GWR model to map the spatial distribution of PM2.5 concentration with fine spatial resolution.

2. The reason why GWR model did not improve $R^2$ significantly may be that the study area was limited to Beijing, the research scope was small, the number of monitoring stations within the research scope was small and the distribution was close, and the effective monitoring data was insufficient in data processing. The process led to weak data differences, and subsequent research will expand the study area to get more effective data and enhance the data difference.
ACKNOWLEDGMENTS

This research was financially supported by Guangxi Key Laboratory of Spatial Information and Geomatics (16-380-25-08), Guangxi young and middle-aged teacher’s basic ability improvement project (KY2016YB192).

REFERENCES


