Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

Improving the Accuracy of $PM_{2.5}$ Concentration Prediction Using Localized Explanatory Factors: A Comparison of GWR, Multiscale GWR, GW Lasso and GW Elastic Net

Miao Fu*
School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, China
*Corresponding Author.

Abstract:

In this paper, $PM_{2.5}$ concentrations are predicted for all counties in China, using the geographically weighted regression (GWR), the geographically weighted Lasso, the geographically weighted Elastic net, and multiscale GWR models. Predictor variables include spatially localized county-level economic activities, population, road network, land use, aerosol optical depth, meteorological and topographic factors. Economic, population, road network and land use data are localize (within 8 Km from the locations studied) to improve the accuracy of the prediction. We found that incorporation of geographic weights into the Lasso and Elastic net models cannot enhance the prediction capacity of them. Multiscale GWR can partially correct the underestimation problem of the GWR model, but presents a lower cross validation R2, and proves to be a time-consuming algorithm. Among those models, GWR is the best model with the highest cross validation R2 (0.8276), and lowest RMSE (7.4752), MAE (5.3904) and MAPE (0.1127). The county-level $PM_{2.5}$ concentration map predicted by GWR is presented.

Keywords*: $PM_{2.5}$ concentrations, GWR, MGWR, Lasso, Elastic net.*

I. INTRODUCTION

About 70% of the population in East Asia are living in an average concentration of fine particulate matters ($PM_{2.5}$) above the WHO interim target of 35 $\mu g/m^3$, 92% of the global population live in areas where concentrations of $PM_{2.5}$ are above 10 $\mu g/m^3$ [1], and the pollutant causes over 3 million annual premature deaths [2]. Thus, accurate prediction of $PM_{2.5}$ concentrations can assist people, especially for outdoor workers, in avoiding exposure to the pollutant. With regard to the prediction of $PM_{2.5}$ concentrations, van Donkelaar et al. (2015) derived 10 km × 10 km global $PM_{2.5}$ estimates from 1998 to 2012 from satellite sources and found that global ambient $PM_{2.5}$ concentrations annually increased 0.55 $\mu g/m^3$, and average $PM_{2.5}$ concentrations in eastern China's urban areas are larger than 80 $\mu g/m^3$ [3]. Their estimates are popularly used in academic studies. The correlation coefficient and the slope between estimated concentrations and ground-based observations outside North America and Europe are 0.81 and

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

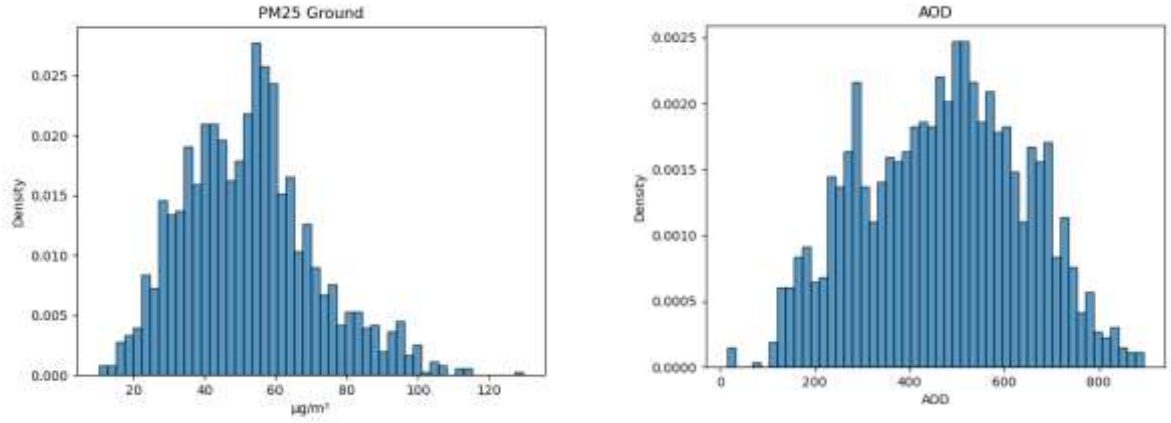
0.68, respectively. The quality of the prediction can be improved if more factors are included. Based on $PM_{2.5}$ composition estimated from a chemical transport model and aerosol optical depth (AOD), van Donkelaar et al. (2019) predicted concentrations of several air pollutants with improved accuracy using a geographically weighted regression over 2000−2016 [4]. Their study focuses on North America. In China, the 11^{th} and 12^{th} Five-Year Plan of China set emission reduction targets for the precursors of $PM_{2.5}$ [5], and hence air quality had significantly improved since then. Therefore, it is necessary to give a new version of the $PM_{2.5}$ concentration distribution which is not only derived from satellite data, but also considers the localized economic activity, population, road network and land use.

Although there are other methods used in the prediction of $PM_{2.5}$ concentrations, such as the dynamic spatial panel model [2], the neural network [6] and the random forest [7], the geographically weighted regression (GWR) remains one of the most frequently used method in $PM_{2.5}$ concentration estimation [4,5,8]. In comparison with existing studies, our study also uses the geographic weighted method mentioned above. However, we focus on county-level data in China, which enables us to use more accurate socioeconomic data within a certain distance from the studied points and other localized transport network, land use, meteorological and topographic factors, to improve the accuracy of prediction. In addition to GWR, geographic weights are also fused with other methods in this study, such as the Lasso model with geographic weight and the Elastic net model with geographic weights. An improved version of GWR, i.e., multiscale GWR is also tested in this paper to find out the best solution within the method group of geographically weighted approaches.

II. DATA

The observed ground-level $PM_{2.5}$ concentrations are obtained from the China Environmental Monitoring Center (CEMC). Annual averages of $PM_{2.5}$ concentrations in 2015 are calculated for each monitoring station. Totally, data from 1494 stations are kept after removing missing values. Similar to Chen et al. (2019) [9], aerosol optical depth (AOD) retrieved from NASA MODIS (Moderate Resolution Imaging Spectroradiometer) are used as the main explanatory variable of $PM_{2.5}$. Especially, within this database, the Multi-angle Implementation of Atmospheric Correction (MAIAC) product (MCD19A2) with a resolution of 1km, which integrates Tera and Aqua images, is chosen as Fu et al. (2020) indicated that it is a good predictor for $PM_{2.5}$ [2]. The distributions of $PM_{2.5}$ and AOD are shown in Figure 1. The quantile distributions of AOD and observed $PM_{2.5}$ concentrations are presented in Table I. The distributions of AOD and ground-level concentrations illustrate a similar pattern, and thus AOD tends to provide an unbiased estimation of $PM_{2.5}$ concentrations in terms of distributions.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021



a. Histogram of ground-level $PM_{2.5}$ concentration

b. Histogram of AOD

Figure 1 Comparison of the distributions of ground-level $PM_{2.5}$ concentrations and AOD

TABLE I. Quantile distributions of ground-level $PM_{2.5}$ concentrations and AOD

	N-Obs	Mean	Std	Min	0.25	0.5	0.75	Max
Ground-level $PM_{2.5}$	1494	52.31	18.49	10.236	38.96	52.06	62.44	129.71
AOD	1494	469.22	171.73	14	338.13	477.90	596.91	895.15

Weather conditions, such as wind and precipitation, affect the diffusion and removal of the pollutants. Meteorological data for county points are sourced from ERA-Interim data, except that relative humidity is from NCEP FNL. After correcting the multicollinearity issue of predictor variables and checking the fitness of the models, among those meteorological variables, 10 meter wind speed (10SI), 2 meter temperature (2T), boundary layer height (BLH), relative humidity (RH), surface pressure (SP), surface solar radiation (SSR) and total precipitation (TP) are used in our model. As plants can partially purify the polluted air, the enhanced vegetation index (EVI) from MODIS is considered in our model.

County-level data, such as GDP, output values and labors of the primary industry, secondary industry and tertiary industry, and industrial emissions are obtained from the statistical yearbooks of the corresponding prefecture cities. Population density is sourced from WorldPop, and has been adjusted with county-level census data. Road network data are from OpenStreetMap, downloaded in Mar. 2018, to make it more consistent with the pollution data in terms of sampling date. Topographic data, such as the Digital Elevation Model (DEM), are from NASA SRTM (Shuttle Radar Topography Mission), and slopes are calculated based on DEM. Land use data are from GlobeLand30.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

III. METHODS

The econometric model of geographically weighted regression is shown in Equation (1). Where y_i is the $PM_{2.5}$ concentration at location i, and β_{0i} and β_{ki} change with geospatial coordinates of location i. $x_{i,k}$ is the explanatory variables that have been described in the previous section. The coefficient vector β_i can be estimated with Equation (2), where X is the matrix of the explanatory variables, W(i) is a diagonal weight matrix that weights in the diagonal are calculated with the bisquare function given in Equation (3). d_{ij} is the distance between location i and j, b means the bandwidth, and $'$ denotes the matrix transpose operation.

$$y_i = \beta_{0i} + \sum_{k=1}^{m} \beta_{ki} x_{i,k} + \varepsilon_i \quad (1)$$

$$\hat{\beta}_i = [X'W(i)X]^{-1}X'W(i)y \quad (2)$$

$$w_j(i) = \begin{cases} (1 - \frac{d_{ij}^2}{b^2})^2, & if\, d_{ij} \le b \\ 0, & if\, d_{ij} > b \end{cases} \quad (3)$$

To improve the prediction accuracy, population, GDP, industrial output values, sums of road network lengths, means of slopes, areas of different types of land covers are all mean or sum values within 8 Km from location i. Sums of road network lengths, means of slopes, areas of different types of land covers are calculated based on GIS. Population within 8 Km can be estimated by summing up the population density within an 8Km circle from location i. GDP and industrial output values are usually provided at a county level, so we allocated them to locations by the proportion of the population within 8 Km in the total population of that county. 8 Km buffer is applied in this study as Zhai et al. (2018) found that land use data acquired within 8 Km have significant effects on $PM_{2.5}$ concentrations [8]. Knibbs et al. (2014) also found that except the satellite data of the pollutant, the next largest contributors of the pollutant columns are roads within 8 Km and industrial land use within 10 Km [10]. Local road network lengths are incorporated in the model as traffic plays an important role in urban air pollution, especially in cities where large point pollution sources are moved outside the city [11].

As multicollinearity issue may exist in explanatory variables, the Least Absolute Shrinkage and Selection Operator (LASSO) [12] and Elastic net [13] approaches are also tried with our data. Chen et al. (2019) suggested using Lasso or Elastic net in fine particles pollution estimation [9]. The Lasso model can reduce the variability of the coefficients by limiting the sum of the absolute values of β_i, and allows for factor selection by shrinking some of the coefficients to zero. In contrast with the previous studies, in our study, geographic weights are combined with the Lasso model (GWLasso), and the coefficient vector β_i of GWLasso can be obtained with Equation (4), where the last term starting with λ is the penalty term, and the main part of this term is the sum of the absolute values of β_i.

$$\hat{\beta}_i^{GWLasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y - \beta_{0i} - \sum_{k=1}^{m} \beta_{ki} x_{ik})^2 + \lambda \sum_{k=1}^{m} |\beta_{ki}| \right\} \quad (4)$$

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

As Lasso tends to remove one variable out of a set of correlated variables, this causes that the estimated coefficients become unstable. To stabilize the model, an l2 penalty term is added and this results in the Elastic net model. After adding the geographic weights, we change the normal Elastic net model into GW Elastic net model, and the estimation of the coefficients is given in Equation (5).

$$\hat{\beta}_i^{GWElasticNet} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y - \beta_{0i} - \sum_{k=1}^{m} \beta_{ki} x_{ik})^2 + \lambda\rho \sum_{k=1}^{m} |\beta_{ki}| + \lambda(1-\rho) \sum_{k=1}^{m} \beta_{ki}^2 \right\} \quad (5)$$

In comparison with GWR, Multiscale GWR (MGWR) assigns different bandwidths for different predictor variables, i.e., varying at different spatial scales [14]. The econometric model of MGWR is given in Equation (6), where subscript b of β indicates the bandwidth used for that particular explanatory variable.

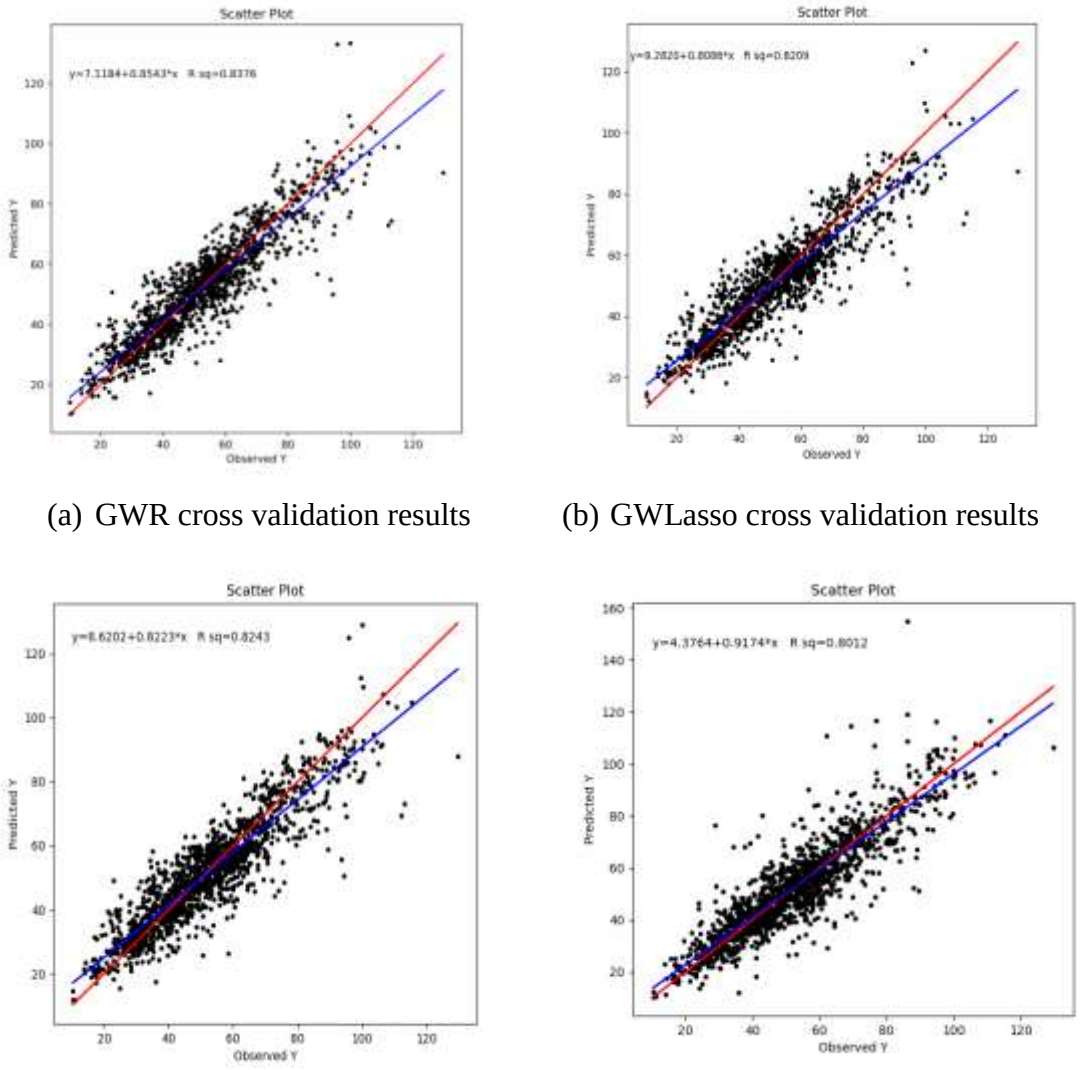
$$y_i = \beta_{b0i} + \sum_{k=1}^{m} \beta_{bki} x_{i,k} + \varepsilon_i \quad (6)$$

After a forward stepwise selection, and by cross validating, the most relevant variables, AOD, localized output values of the secondary industry(Ind2_8K), localized sums of the road network lengths (Roadnet_8K), localized population (Pop_8K), localized mean slopes (Slope_8K), 10SI, 2T, BLH, RH, SP, SSR, TP, DEM and EVI, are kept in the model. The suffix 8K in variable names indicates that the values are localized within 8Km from the location, and variables without the 8K suffix use point values exactly sampled at the spatial location.

IV. RESULTS

To assess the quality of the prediction results from models mentioned about, the leave-one-out cross validation (CV) is used for the GWR, GWLasso, GW Elastic Net models, and 80-fold cross validation is used for the MGWR model. As geographically weighted regressions highly depend on the spatial weight matrix, the 10-fold cross validation is not suitable for our case as it reduces the dimensions of the spatial weight matrix by 1/10, which can significantly distort the results. Thus, the leave-one-out cross validation is the best choice in our study. However, running MGWR for a leave-one-out cross validation may take several months. To keep it in an acceptable time span, an 80-fold cross validation is used for MGWR instead, which does not distort the results much (checked with experiments). The cross validation results of the GWR, GWLasso, GW Elastic net and MGWR models are shown in Figure 2 (a), (b), (c) and (d), respectively.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021



(a) GWR cross validation results (b) GWLasso cross validation results

(c) GW Elastic net cross validation results (d) MGWR cross validation results

Figure 2 Cross validation results of GWR, GWLasso, GWElastic net and MGWR

From Figure 2, one can see that GWR has the highest cross validation R^2 of 0.8376 and the second large slope of 0.8543, which means that it fits the sample best. However, GWR slightly underestimates the concentrations as the slope is lower than one. MGWR, as a result of using variable bandwidths, performs better in terms of the slope, but does not fit the sample as well as GWR. Other detailed cross validation outcomes are given in Table II, including RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) and correlation coefficients, which suggests that GWR is the best model, with the lowest RMSE, MAE and MAPE among those four models. Therefore, predicted $PM_{2.5}$ concentrations for counties are calculated with the GWR model and are presented in Figure 3. The first row of the table presents the correlation between AOD and the ground-level concentrations, suggesting that all models significantly improve the prediction capacity of AOD by adding more predictor factors and

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

using advanced models.

TABLE II. Comparison of the cross validation results

Model	r	R^2	Slope	RMSE	MAE	MAPE
AOD	0.4971	0.2471				
GWR	0.9152	0.8376	0.8543	7.4752	5.3904	0.1127
GWLasso	0.9061	0.8209	0.8086	7.8569	5.7748	0.1208
GW Elastic Net	0.9078	0.8243	0.8223	7.7792	5.6781	0.1190
MGWR	0.8951	0.8012	0.9174	8.5853	5.9687	0.1214

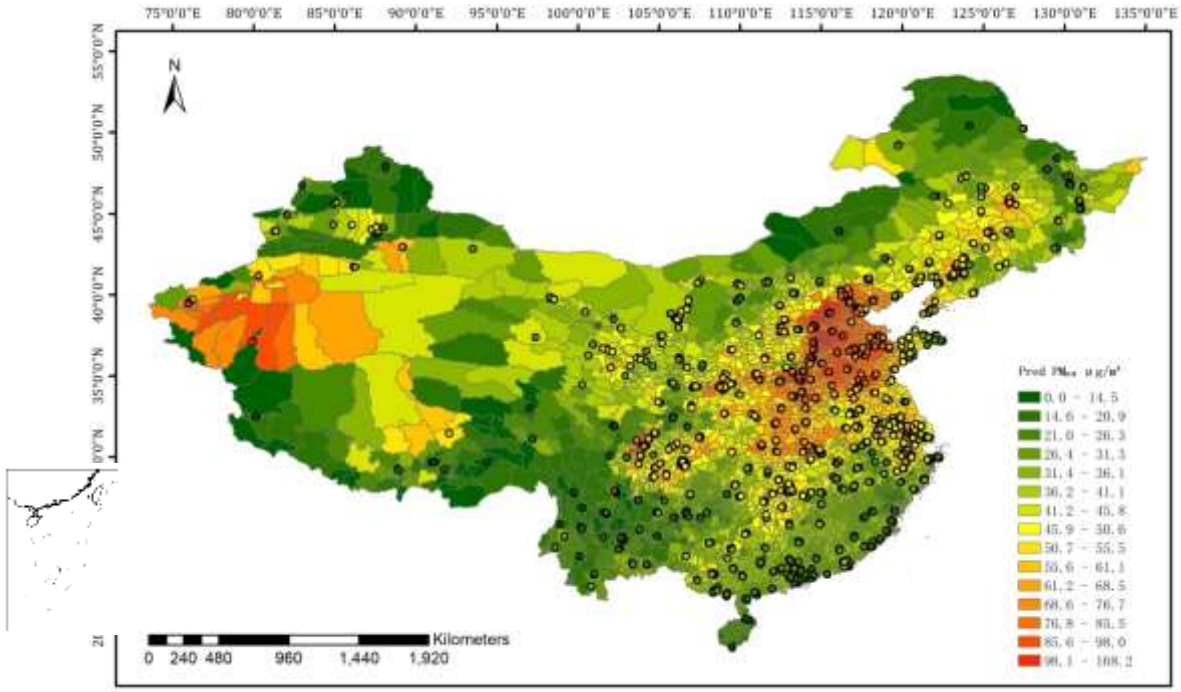


Figure 3 County-level $PM_{2.5}$ concentrations predicted by GWR

Circle points in Figure 3 indicate the locations of the monitoring stations, and the color in the polygons shows the predicted concentrations of counties. Figure 3 demonstrated high agreement between the monitoring values and the predicted concentrations. The model also presents predicted concentrations for those counties without monitoring stations. It is noteworthy that as the locations of the urban areas of the counties are used in our study, the concentrations shown in the map only represent the central cities of the counties, not the rural areas. If locations of rural areas are used to collect the explanatory variables, concentrations of the rural areas can also be predicted.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

Compared with some existing studies, our study presents predicted concentrations with relatively higher accuracy. For example, the cross validation correlation coefficient of van Donkelaar et al. (2015) is 0.81 [3], in contrast with r = 0.9152 in our study. The cross validation R^2 from the GWR model of Zhai et al. (2018) is 0.831 [8], and cross validation R^2 = 0.70 from Van Donkelaar et al. (2019) [4], in comparison with our cross validation R^2 of 0.8376. Cross validation RMSE from our prediction is 7.4752 μg/ m^3, which is lower than 9.3 μg/ m^3 from Xiao et al. (2020) [5].

V. CONCLUSION

The comparison among GWR, GWLasso, GW Elastic net and MGWR implies that the complicated models, such as the Lasso and Elastic net models do not perform better than the GWR model, in terms of $PM_{2.5}$ concentration prediction. This suggests that the limits set to the coefficients of the regressions may cause more troubles than they resolve, and probably multicollinearity is not so serious within the explanatory variables after the variable selection. The Multiscale GWR model, i.e., the multi-bandwidth GWR model, can partially correct the underestimation of the GWR model. However, it also reduces the accuracy of the prediction as the quality of its prediction highly depends on the bandwidths used. MGWR proves to be a very time-consuming algorithm. The findings above are only relevant to $PM_{2.5}$ prediction. In our research, we found that the appropriate method may vary with the pollutant studied.

Our concentration prediction, which considers localized socioeconomic, geographic and meteorological factors, presents more accurate estimates and covers areas without monitoring stations. The predicted results suggest that $PM_{2.5}$ concentrations are low in the mountainous areas of the county, low in the rainy areas, such as the Southern China, and high in the Northern China regions with high intensity of economic activities and high density of population. It is clear that the air quality of China had been significantly improved since the implementation of various clean air policies.

ACKNOWLEDGEMENTS

The work on this paper has been supported by Humanities and Social Science Foundation of the Ministry of Education of China (Estimation of spatially distributed marginal health damage values of $PM_{2.5}$ pollution in China, 17YJA790021), Natural Science Foundation of Guangdong Province (2017A030313439) and special fund from Institute of City Strategy Studies, Guangdong University of Foreign Studies (JDZB202108). The views expressed are those of the authors.

REFERENCE

[1] World Health Organization (2018) WHO Global Ambient Air Quality Database; WHO: Geneva.
[2] Miao, F, Kelly, A., & Clinch, P. (2020) Prediction of $PM_{2.5}$ daily concentrations for grid points throughout a vast area using remote sensing data and an improved dynamic spatial panel model. Atmospheric Environment, 237: 117667.
[3] Van Donkelaar A, Martin RV, Brauer M, Boys BL (2015) Use of satellite observations for long-term exposure

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 291-299
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

assessment of global concentrations of fine particulate matter. Environ Health Perspect 123: 135–143.

[4] Van Donkelaar, A., Martin, R. V., Li, C., & Burnett, R. T. (2019) Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. Environmental Science and Technology, 53(5): 2595-2611.

[5] Xiao, Q., Geng, G., Liang, F., Wang, X., Lv, Z., Lei, Y., Huang, X., Zhang, Q., Liu, Y. and He, K. (2020) Changes in spatial patterns of PM2.5 pollution in China 2000–2018: Impact of clean air policies. Environment International.2020, 141: 105776.

[6] Wang, W., Zhao, S., Jiao, L., Taylor, M., Zhang, B., Xu, G., & Hou, H. (2019) Estimation of PM2. 5 Concentrations in China using a spatial back propagation neural network. Scientific reports, 9(1), 1-10.

[7] Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., De'Donato, F., et al. (2019) Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment international, 124, 170-179.

[8] Zhai, L., Li, S., Zou, B., Sang, H., Fang, X., & Xu, S. (2018) An improved geographically weighted regression model for pm 2.5 concentration estimation in large areas. Atmospheric Environment, 181(MAY): 145-154.

[9] Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketrzel, M. et al. (2019) A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. Environment International, 130: 104934.

[10] Knibbs, L.D., et al. (2014) A national satellite-based land-use regression model for air pollution exposure assessment in Australia. Environ. Res. 135, 204–211.

[11] Health Effects Institute (2010) Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Health Effects Institute: Boston, Special Report 17.

[12] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society B 58(1): 267-288.

[13] Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 301–320. MR2137327.

[14] Yang, W. (2014) An extension of geographically weighted regression with flexible bandwidths. PhD thesis, School of Geography and Geosciences, University of St. Andrews, Fife, Scotland, UK. http://hdl.handle.net/10023/7052.