Sensors and Materials, Vol. 32, No. 12 (2020) 4393–4412
MYU Tokyo

S & M 2417

Optimization of Seasonal Geographically and Temporally Weighted Regression Model for Accurate Estimation of Seasonal $PM_{2.5}$ Concentrations in Beijing–Tianjin–Hebei Region

Lei Zhou,[1] Yani Wang,[2*] Mingyi Du,[1] Changfeng Jing,[1] Siyu Wang,[1]
Yinuo Zhu,[1] Ting Luo,[1] Congcong He,[1] Ting Gao,[1] and Kun Yang[1]

[1]School of Geomatics and Urban Spatial Informatics,
Beijing University of Civil Engineering and Architecture, Beijing 102616, China
[2]Gansu Institute of Natural Resources Planning and Research, Gansu 730030, China

(Received August 28, 2020; accepted December 2, 2020)

Keywords: Beijing–Tianjin–Hebei urban agglomeration, $PM_{2.5}$, S-GTWR, greedy algorithm, model optimization

Particulate matter with a diameter of less than 2.5 μm ($PM_{2.5}$) has a significant impact on air pollution, atmospheric visibility, and human health. The most basic and important step of regional air pollution control is to obtain air pollution data in different seasons from both satellite sensors and ground-level observations. The aim of this paper is to accurately estimate the $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei urban area in different seasons by establishing a seasonal geographically and temporally weighted regression (S-GTWR) model that integrates multiple complex factors. Using a greedy algorithm, the model results were optimized by selecting the characteristic variables that contributed to the accuracy of the model in different seasons. The measured and estimated $PM_{2.5}$ concentrations were compared and the cross-validation results were used as a basis for evaluating the accuracy of the model. The results showed that the accuracy of the S-GTWR model that combined the optimal characteristic variables was higher than that of the geographically weighted regression (GWR) model and the kriging method. The mean prediction error (ME), relative prediction error (RPE), and root mean square error (RMSE) of the S-GTWR model were small, and the coefficient of determination (R^2) of the model exceeded 0.86 for each season. The accuracy of the S-GTWR model in estimating the $PM_{2.5}$ concentration was highest in summer and lowest in winter. In addition, the proposed model can accurately estimate $PM_{2.5}$ concentrations in areas without monitoring sites. The results can provide a scientific basis for the study of pollution control and $PM_{2.5}$ exposure in large urban agglomerations.

1. Introduction

In recent years, with the acceleration of China's social and economic development, urban integration has developed rapidly and industry has expanded similarly.(1) Subsequently, the

*Corresponding author: e-mail: yani_wang0504@126.com
https://doi.org/10.18494/SAM.2020.3077

ISSN 0914-4935
https://myukk.org/

problem of environmental pollution caused by the increases in population and industrial production has attracted increasing public attention.[2] China's three economic heartlands, the Yangtze River Delta, the Pearl River Delta, and the Beijing–Tianjin–Hebei region, are particularly seriously impacted by air pollution caused by economic development.[3,4] Particulate matter 2.5 ($PM_{2.5}$), also known as fine particulate matter, refers to particulate matter in the atmosphere with an aerodynamic diameter smaller than 2.5 μm.[5] $PM_{2.5}$ can remain suspended in the air for a long time, and it is difficult for it to disperse in a short time once it accumulates. The adsorption of $PM_{2.5}$ greatly increases the risk of disease for urban residents.[6] In addition, derivatives of air pollutants also cause global climate change, acid rain, and other environmental problems.[7] The accurate estimation of near-surface $PM_{2.5}$ concentrations and the analysis of the spatial differences and variation characteristics of $PM_{2.5}$ concentrations would greatly contribute to solving environmental problems.

A number of studies have shown that spatially continuous $PM_{2.5}$ concentrations can be obtained through global or regional chemical transport models. These models require detailed pollutant emission inventories as the base data and establish chemical transformation models for air pollutants under a series of ideal assumptions. Therefore, numerical simulation methods provide uncertain results in the study of the spatially continuous distribution of $PM_{2.5}$ concentrations.[8] Non-mechanistic models, which are represented by mathematical statistics and machine learning models, are widely applied to pollutant concentration predictions. Machine learning is a new method in the field of artificial intelligence. Through the effective learning of the characteristics of a large number of observation data, machine learning provides new research ideas and methods for the study of the spatially continuous distribution of the $PM_{2.5}$ concentration.[9] Artificial neural network models based on meteorological and air quality data have exhibited good performance in the study of the daily average $PM_{2.5}$ concentration.[10] Although machine learning methods have been used to improve the fitting accuracy in the study of pollutant trends, they also have the following problems. First, the data used in machine learning models are mainly based on the real-time data of air pollutants obtained from ground monitoring stations.[11] However, most of the existing monitoring stations in China are located in urban areas. It is very difficult to obtain the real-time concentration of pollutants in suburban and rural areas without monitoring stations. The uneven distribution of data will lead to a reduction in model accuracy. Then, the machine learning method will cause problems such as overfitting and local minima, resulting in an insufficient generalization ability.[12]

A statistical method developed in the middle of the 20th century was first applied to the study of air pollutants. The commonly used kriging method was proposed by Krige in 1960. Menz *et al.* formalized the method of giving it the ability to address geostatistical problems.[13] However, owing to the limited number and uneven distribution of monitoring stations on the ground, the kriging method cannot accurately estimate the $PM_{2.5}$ concentration in areas without monitoring stations. Moreover, it does not consider the impact of time and space dimensions on air pollutants, making it difficult to meet the requirements for the study of the spatiotemporal distribution of the $PM_{2.5}$ concentration.[14] A large number of studies have confirmed that regression analysis can also be used to estimate the $PM_{2.5}$ concentration. The geographically weighted regression (GWR) model has been used in research on air pollutants.[15] The GWR

model is better than the ordinary linear regression model in revealing the spatial heterogeneity of the $PM_{2.5}$ concentration, but the fluctuation in the estimated $PM_{2.5}$ concentration over time is ignored in the modeling process.[16,17] To apply the local regression model to the analysis and research of spatiotemporal data, Huang *et al*. added the time factor to the GWR model and established the geographically and temporally weighted regression (GTWR) model. This model has been suggested to have a better fitting effect when verifying spatiotemporal data.[18] Domestic scholars have studied the extension and application of the GTWR model. They found that the precision of the same GTWR model differed in different regions and at different times, so the GTWR model needs to be improved.

The selection of characteristic variables to be used in the regression model is a key step to improve model performance and accuracy. The observations from satellite-based remote sensors are an important data source for estimating the $PM_{2.5}$ concentration. Aerosol optical depth (AOD) has a strong correlation with the $PM_{2.5}$ concentration owing to its good spatial continuity, so it has been used for estimating the $PM_{2.5}$ concentration on a regional scale.[19] Mirzaei *et al*. used the GTWR model to study the temporal and spatial variability between the $PM_{2.5}$ concentration measured by a ground monitoring station and the satellite AOD data. Meteorological variables and land-use information were used as additional predictors in the GTWR model to improve its accuracy.[20] Fu and Li added social and economic indicators, such as per capita GDP and urban population ratio, and verified the potential relationship between the social and economic indicators and the $PM_{2.5}$ level when using the GTWR model to estimate the $PM_{2.5}$ level worldwide.[21] The GTWR model proposed by He and Huang added some variables related to meteorology and land cover when estimating the $PM_{2.5}$ concentration, and the cross-validation coefficient of determination (R^2) reached 0.80.[22] Guo *et al*. found that the GTWR model with meteorological parameters and land use variables set as predictors could be used to fit the seasonal $PM_{2.5}$ concentration, and a seasonal GTWR (S-GTWR) fitting model could be constructed to obtain the $PM_{2.5}$ concentration in an urban area with high precision.[23]

Through the analysis of existing research, we found that the selection of characteristic variables for the GTWR model mainly depends on the correlation between characteristic variables and observation data. However, in the existing research, the differences in the precision of the same model constructed in different regions and at different times have been ignored. That is, according to the characteristics of the seasonal variation in $PM_{2.5}$ concentration, the optimal characteristic variables of the $PM_{2.5}$ concentration in different seasons should be selected for modeling instead of using the same characteristic variables for different times of the year.[24] Moreover, the previous characteristic variable selection method based on correlation depends on only the degree of correlation between a single characteristic variable and the $PM_{2.5}$ concentration, rather than the degree of the contribution of the characteristic variable in the modeling process as a selection standard for the characteristic variable. Therefore, a characteristic variable with a strong correlation obtained during the process of characteristic variable selection is not necessarily the optimal characteristic variable that can optimize the model accuracy. In addition, the time span of the $PM_{2.5}$ concentration in the existing research is mostly a short time series of one or two years. As a result, the possibility of the reduced model accuracy caused by the effect of extreme weather in the

research time period increases. In particular, when building $PM_{2.5}$ concentration estimation models for different seasons, the seasonal $PM_{2.5}$ concentration over a short time series is not representative.

In view of the problems in the existing research, our research objective is the accurate estimation of the $PM_{2.5}$ concentration (all concentrations are mass concentrations in this study) in the Beijing–Tianjin–Hebei urban agglomeration. In this research, a greedy algorithm is used to select the optimal characteristic variables, and the degree of contribution of characteristic variables in the modeling process is used to screen the characteristic variables. The factors sensitively influencing the $PM_{2.5}$ concentration are integrated to optimize the precision of an S-GTWR model, and the precision of the model is quantitatively evaluated. An important objective of this study is to build the best model for estimating the $PM_{2.5}$ concentration that considers the optimal characteristic variable group, and this model is used to calculate the $PM_{2.5}$ concentration in each season and its spatial distribution in the research area to provide a scientific basis for $PM_{2.5}$ monitoring and control.

2. Materials and Methods

2.1 Study area

The Beijing–Tianjin–Hebei region is the heart of China's Capital Economy Circle. As shown in Fig. 1, this region consists of two municipalities (Beijing and Tianjin) and 11 prefecture-level cities in Hebei Province (Shijiazhuang, Baoding, Langfang, Tangshan, Zhangjiakou, Chengde,

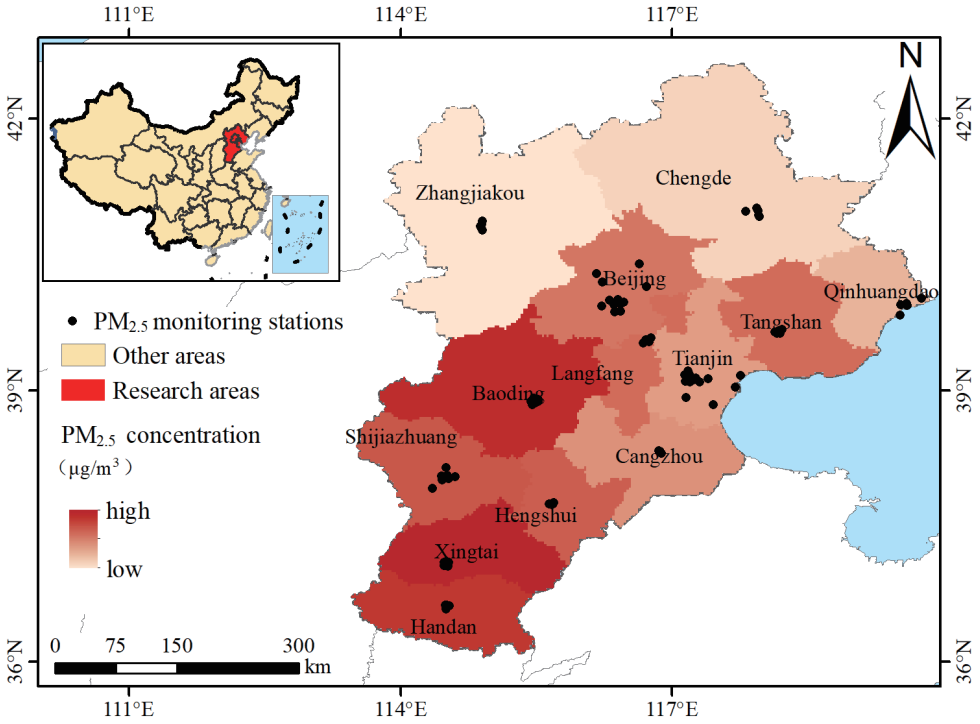

Fig. 1. (Color online) Schematic diagram of study area and $PM_{2.5}$ monitoring stations.

Qinhuangdao, Cangzhou, Hengshui, Xingtai, and Handan). The Beijing–Tianjin–Hebei region is located in the northern part of the North China Plain. The overall altitude is high in the northwest and low in the southeast. The mountains in the northwest are not conducive to the diffusion of pollutants. There are four main types of landform in the region, plateau, mountain, basin, and plain, with various other landform types. In other words, owing to the geographical location and landforms of the Beijing–Tianjin–Hebei region, the concentration of air pollutants sharply increases when conditions are unfavorable for air diffusion, which is likely to cause haze and affect the health and lives of urban residents.[25] The Beijing–Tianjin–Hebei region has a large population (approximately 110 million, or 8% of China's population) and its urbanization has been rapid. This region also suffers from serious air pollution caused by industry, traffic, and coal for heating in winter. During the study period, only Zhangjiakou and Chengde among the 13 cities in the Beijing–Tianjin–Hebei region met China's national annual standard (35 μg/m^3). The $PM_{2.5}$ average in some cities was 300% higher than the national standard and 600% higher than the World Health Organization standard (15 μg/m^3).[26]

The ground monitoring data of the $PM_{2.5}$ concentration in the study were obtained from the environmental monitoring stations of China. $PM_{2.5}$ monthly mean concentrations were calculated by obtaining $PM_{2.5}$ hourly mean concentrations at 79 monitoring sites from 2015 to 2019. The meteorological data were obtained from the Chinese meteorological data network. These data mainly include the daily mean data of air pressure (PRE), air temperature (TEM), precipitation (PRS), relative humidity (RHU), wind direction and wind speed (WIN), sunshine duration (SSD), and ground temperature at 0 cm (GST). Since the numbers of meteorological data and $PM_{2.5}$ monitoring stations are different and inconsistent in terms of spatial location, kriging interpolation was performed on the meteorological data, and the monthly averages of the meteorological data were obtained through resampling for subsequent calculation. MOD/MYD04_3K (full MODIS Terra/Aqua Aerosol 5-min L2 Swath 3 km) data were provided by NASA for Level 2 aerosol products mounted on Terra sensors. The spatial resolution of this dataset is 3 km. Data on topography, population, and land use types were provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences.

2.2 Methodology

2.2.1 S-GTWR model

The S-GTWR model is a set of GTWR models for different seasons. GTWR is an extension of GWR with temporal variations and incorporates both spatial and temporal heterogeneity in the data.[27] The temporal and spatial heterogeneity of $PM_{2.5}$ concentrations is fully considered in the S-GTWR modeling process.[28] At the heart of the GTWR model are temporal and spatial weight matrices. The 3D coordinates (x, y, t) of observation i and other observations are used to construct the weight matrix. GTWR models can be formulated as

$$\begin{aligned} Y_i = {} & \beta_0(u_i, v_i, t_i) + \beta_1(u_i, v_i, t_i) X_{1(i)} + \beta_2(u_i, v_i, t_i) X_{2(i)} + \beta_3(u_i, v_i, t_i) X_{3(i)} \\ & + \beta_4(u_i, v_i, t_i) X_{4(i)} + \cdots + \beta_n(u_i, v_i, t_i) X_{n(i)} + \varepsilon_i . \end{aligned} \quad (1)$$

In the formula, (u_i, v_i, t_i) are the geographical location and time coordinates of the ith sample point. $\beta_0(u_i, v_i, t_i)$ is the intercept value. $\beta_0(u_i, v_i, t_i)$, $\beta_1(u_i, v_i, t_i)$, $\beta_2(u_i, v_i, t_i)$, $\beta_3(u_i, v_i, t_i)$, …, $\beta_n(u_i, v_i, t_i)$ are the coefficients of each characteristic variable involved in modeling, and n is the number of characteristic variables involved in modeling. ε_i is the random error at position i, which obeys $\varepsilon_i \sim N(0, \delta^2)$.

Similarly to that of the GWR model, the regression coefficient $\widehat{\beta_n}(u_i, v_i, t_i)$ of the GTWR model can be expressed as a matrix:

$$\hat{\beta}(u_i, v_i, t_i) = (X^T W(u_i, v_i, t_i) X)^{-1} X^T W(u_i, v_i, t_i) Y, \tag{2}$$

where $W(u_i, v_i, t_i)$ is the weight matrix, which can be expressed as

$$W(u_i, v_i, t_i) = \begin{bmatrix} w_1(u_i, v_i, t_i) & 0 & \cdots & 0 \\ 0 & w_2(u_i, v_i, t_i) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n(u_i, v_i, t_i) \end{bmatrix}. \tag{3}$$

In the formula, $w_1(u_i, v_i, t_i)$, $w_2(u_i, v_i, t_i)$, …, $w_n(u_i, v_i, t_i)$ represent the monotonic decay functions of the space-time distance between fitting point i and other modeling points j. The determination of spatiotemporal proximity is crucial for the calculation of spatiotemporal weights.

2.2.2 Screening methods for characteristic variables

The regression model mainly relies on the correlation between characteristic variables and ground observation sites to estimate the $PM_{2.5}$ concentrations in areas without monitoring sites. The characteristic variables are chosen by selecting the variables with a significant effect on the $PM_{2.5}$ concentration from multiple characteristic variables, and the variables without a significant effect on the accuracy of the model are eliminated. In this paper, a greedy algorithm was used to select the optimal characteristic variables for different seasons and to establish regression models with high reliability. The greedy algorithm is an effective method of obtaining the global optimal or approximate global optimal solution by using the local optimal solution.[29] The principle of the greedy algorithm is to search variables one by one from the set of characteristic variables. An evaluation criterion is used to determine the degree of contribution of the variables to the accuracy of the model, and it is used as the basis for the selection of variables.[30] Cross-approximate entropy (Cross-ApEn) is used as the evaluation basis for the selection of characteristic variables. Cross-ApEn is an index used to determine the similarity degree of two sets of sequences by calculating the conditional probability that they have the same pattern. Its formula is expressed as

$$H_{Cross-ApEn}(m,r) = -\frac{1}{N-m}\sum_{k=1}^{N-m} \ln P_K(B \mid A), \tag{4}$$

where m is the pattern dimension, r is the similarity tolerance, N is the length of the sequence, both A and B represent the similarity of the sample r values, and $P_K(B|A)$ is the similarity probability of r.

First, the regression model of all the characteristic variables is established to obtain the initial Cross-ApEn. One characteristic variable is removed from all the characteristic variables, and the remaining characteristic variables are used to establish the regression model and obtain the Cross-ApEn of the group model (the number of regression models established is n if the number of characteristic variables is n). The model with the highest Cross-ApEn is selected from these regression models. A key feature of this method is that the characteristic variables can no longer be added to the model if they are eliminated.

2.2.3 Accuracy evaluation

Cross-validation was used to evaluate the accuracy of the estimation. The observed $PM_{2.5}$ concentration was used to validate the results of downscaling based on R^2, mean prediction error (MPE), relative prediction error (RPE), and root mean square error (RMSE).[31] All monitoring site data were randomly divided into 10 parts in the modeling process; nine parts were used for modeling and one part was used to validate the results. The corresponding evaluation indicator formulas are as follows:

$$MPE = \frac{1}{N}\sum_{i=1}^{N}[Z(x_i, y_i) - Z^*(x_i, y_i)], \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Z(x_i, y_i) - Z^*(x_i, y_i)]}, \tag{6}$$

$$PRE = \frac{RMSE}{\hat{y}} \times 100\%, \tag{7}$$

where N is the number of monitoring points, $Z(x_i, y_i)$ is the measured value of inspection point i, $Z^*(x_i, y_i)$ is the predicted value of inspection point i, and (x_i, y_i) are the position coordinates of inspection point i.

The MPE reflects the closeness of the measured and predicted values. RMSE is a measure of modeling accuracy. Using the valuation sensitivity and extreme value effects of the sample data, the optimal modeling method was determined by the RMSE.[32] A lower MPE and a lower RMSE represent a higher prediction accuracy and a lower prediction bias.

3. Results and Discussion

3.1 Modeling accuracy

A large number of observations and research results of atmospheric pollutants showed that the mass concentration of air pollutants is not only related to pollution sources but also affected by social, economic, and natural factors in the region. The prediction performance of the $PM_{2.5}$ concentration is also different in different time periods and regions, so the selection of characteristic variables directly affects the accuracy of the model. Figure 2 shows the Cross-ApEn plotted against the number of iterations during the selection of characteristic variables for modeling each season. In the figure, • represents the global initial mutual approximate entropy. * is the model with the minimum mutual approximate entropy in each iteration, and the corresponding variable is the eliminated variable. × is the iteration termination, and the iteration termination condition is that the Cross-ApEn of this iteration is less than or equal to that in the previous iteration.

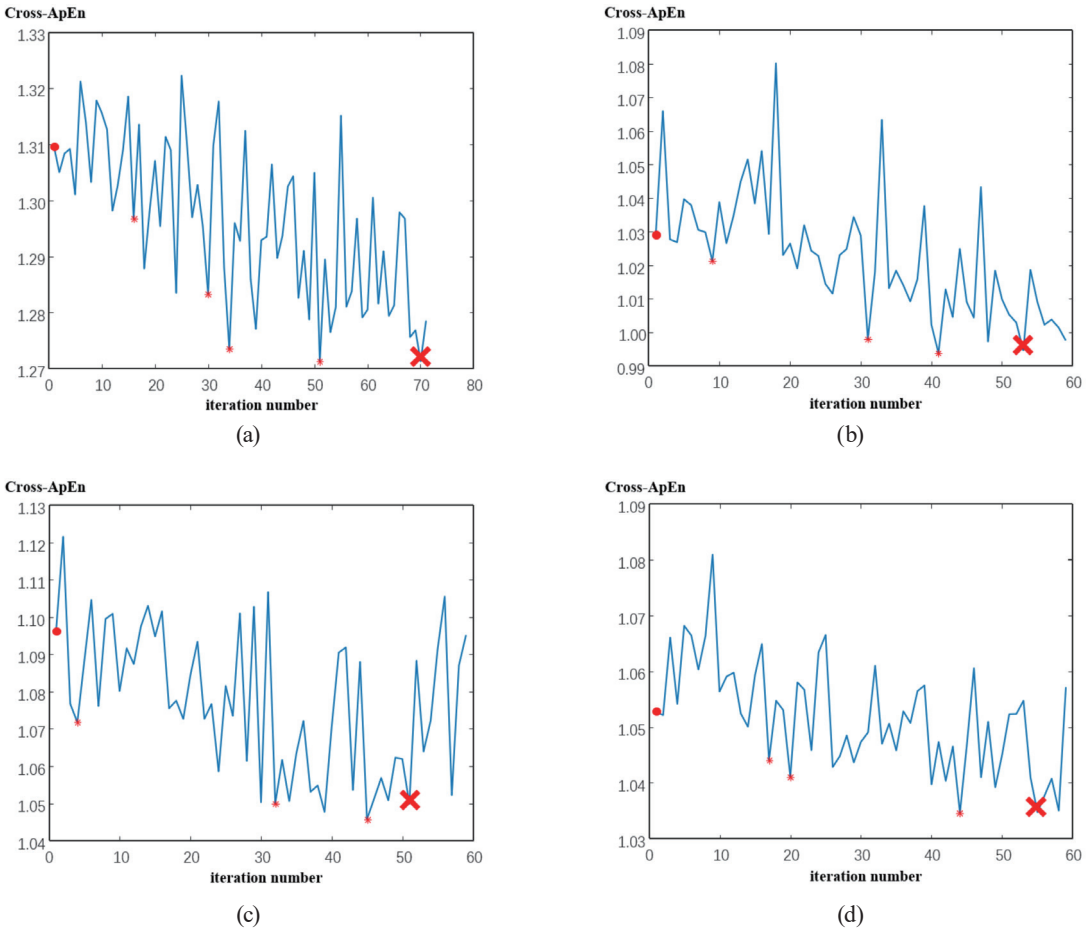

Fig. 2. (Color online) Mutual approximate entropy plotted against number of iterations for each season: (a) spring, (b) summer, (c) autumn, and (d) winter.

Up to 10 characteristic variables could be selected in this experiment. Figure 2(a) shows that after 59 iterations, four characteristic variables were eliminated in the spring $PM_{2.5}$ concentration estimation model. The summer model [Fig. 2(b)] was iterated 46 times and three characteristic variables were eliminated. The autumn model [Fig. 2(c)] was iterated 46 times and three characteristic variables were eliminated. The winter model [Fig. 2(d)] went through 46 iterations and three characteristic variables were eliminated. The characteristic variables finally selected for each seasonal model are shown in Table 1.

The S-GTWR models of different seasons in Beijing–Tianjin–Hebei were constructed by using the multiple variables selected for different seasons as the auxiliary variables in the modeling process. To clarify the difference between the $PM_{2.5}$ concentrations estimated by different seasonal models and the measured concentrations, the difference between the measured and estimated concentrations was calculated, and the results were displayed spatially. Figure 3 shows the difference between the observed and estimated $PM_{2.5}$ concentrations in different seasons.

By comparing the estimated $PM_{2.5}$ concentrations of the S-GTWR model with the measured $PM_{2.5}$ concentrations of ground-based monitoring stations in different seasons, it is obvious that there is a large gap between the estimated and measured $PM_{2.5}$ concentrations for the winter model. Considering that winter is cold in North China, coal consumption for heating increases during this period, which leads to an increase in $PM_{2.5}$ concentration. Because of the effects of human factors on $PM_{2.5}$, the precision of the winter model is slightly lower than that of the other models. The $PM_{2.5}$ concentration in spring is affected not only by natural and social economic factors but also by the lag of the $PM_{2.5}$ concentration in winter, which makes it difficult to reduce the $PM_{2.5}$ concentration in spring in a short time; therefore, the difference between the estimated $PM_{2.5}$ concentration based on the spring model and the observed $PM_{2.5}$ concentration in some areas is greater than that in winter. In contrast, the difference between the $PM_{2.5}$ concentrations estimated by the summer and autumn models and those measured at monitoring sites is relatively small.

To evaluate the $PM_{2.5}$ concentration estimated by the model, the cross-validation method is used to test the model accuracy. At the same time, to further understand the effect of spatiotemporal information on the performance of the $PM_{2.5}$ concentration estimation model, in this paper, the kriging interpolation algorithm and GWR model with the same independent variables as the S-GTWR model are added to simulate the $PM_{2.5}$ concentration in different seasons in the Beijing–Tianjin–Hebei region. The results of the cross-validation of the three models are shown in Table 2.

Table 1
Results of variable selection for different seasons.

Season	Characteristic variables in S-GTWR						
Spring	AOD	GST	SSD	TEM	WIN	PX	
Summer	AOD	GST	PRE	RHU	SSD	WIN	LC
Autumn	AOD	PRS	RHU	SSD	TEM	WIN	LC
Winter	AOD	GST	PRS	RHU	SSD	TEM	WIN

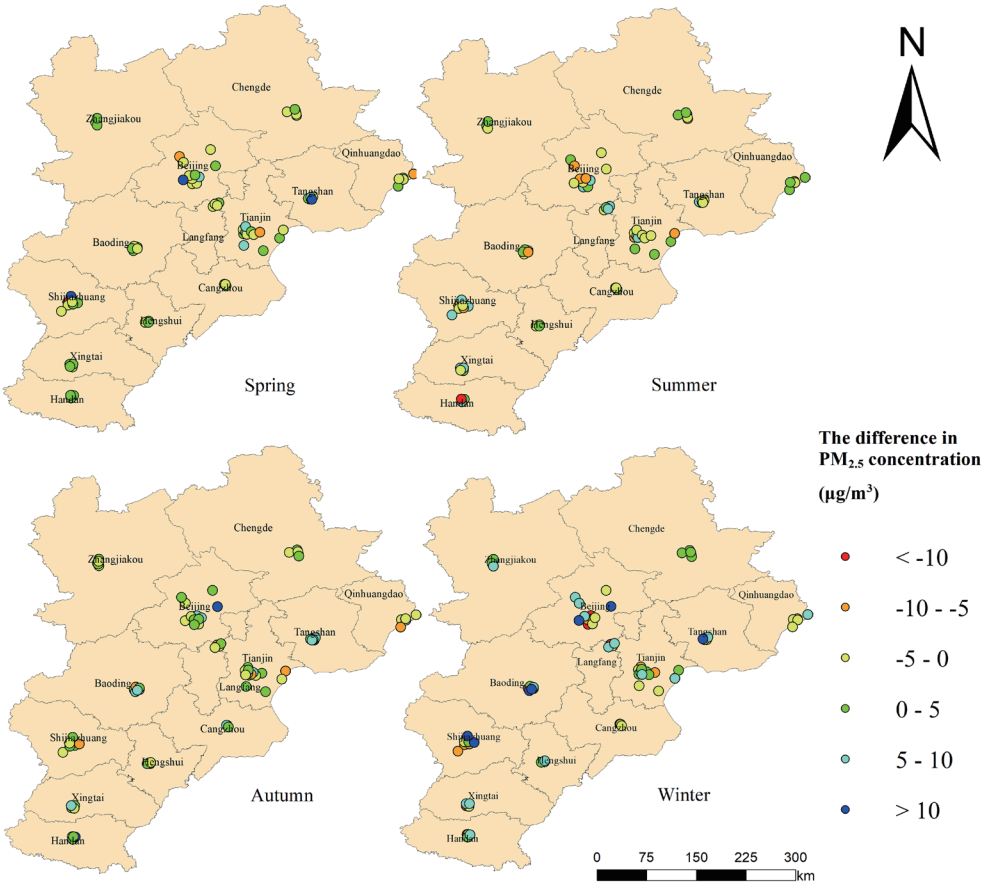

Fig. 3. (Color online) Difference between observed and estimated $PM_{2.5}$ concentrations in different seasons.

Table 2
Comparison results of the accuracy.

Season/ method	S-GTWR			GWR			Kriging		
	MPE	RMSE	RPE (%)	MPE	RMSE	RPE (%)	MPE	RMSE	RPE (%)
Spring	0.88	8.13	7.51	1.32	10.71	10.01	1.78	11.67	14.15
Summer	0.67	7.50	8.91	0.81	9.48	9.51	1.23	11.17	12.78
Autumn	0.96	11.20	10.01	1.21	13.57	13.33	1.94	23.50	24.48
Winter	1.36	9.46	10.75	1.53	16.51	16.29	1.89	20.66	20.89

Note: MPE and RMSE units are $\mu g/m^3$.

By comparing and analyzing the related index results, it is concluded that (1) the kriging interpolation method without considering the spatiotemporal variation in $PM_{2.5}$ exhibits the largest MPE and RMSE values, and the model precision is low; the estimated concentration of $PM_{2.5}$ differs greatly from the observed concentration at the ground monitoring stations. (2) The results of the seasonal fitting of the $PM_{2.5}$ concentration obtained by the GWR model considering the geographical location are better than those obtained by kriging, but the results of the fitting of the $PM_{2.5}$ concentration are comparable for all seasons. The S-GTWR model, which considers both time and space factors, is the best among all the models. The results obtained by the S-GTWR model are superior to those obtained by all the other methods. The

decrease in MPE ranged from 20 to 40% μg/m^3, the decrease in RMSE ranged from 30 to 50% μg/m^3, and the decrease in RPE ranged from 3.87 to 11.77%. (3) The comparison of the estimated $PM_{2.5}$ concentrations in different seasons indicated that the performance of the seasonal model in spring and summer is better than that in autumn and winter. This difference is mainly caused by the atmosphere being generally stable over a short time span in cooler seasons, which means that $PM_{2.5}$ estimates are heavily dependent on spatial emissions; in warmer seasons, since changes in weather patterns greatly affect the formation and dispersion of air pollutants, the estimation of $PM_{2.5}$ mainly depends on meteorological factors. The degree of precision of the optimized model in autumn and winter is obviously higher than that in spring and summer when spatial and time information is added. The results show that the spatial variability information incorporated into the S-GTWR model plays an important role in the cold season, while the temporal variability information plays an important role in summer.

From the results of the analysis of the existing indices, the R^2 values of S-GTWR, GWR, and kriging were calculated, and scatter plots were drawn (Figs. 4–6). This plot was used to analyze the relationship between the measured and estimated $PM_{2.5}$ concentrations for different seasons for each method.

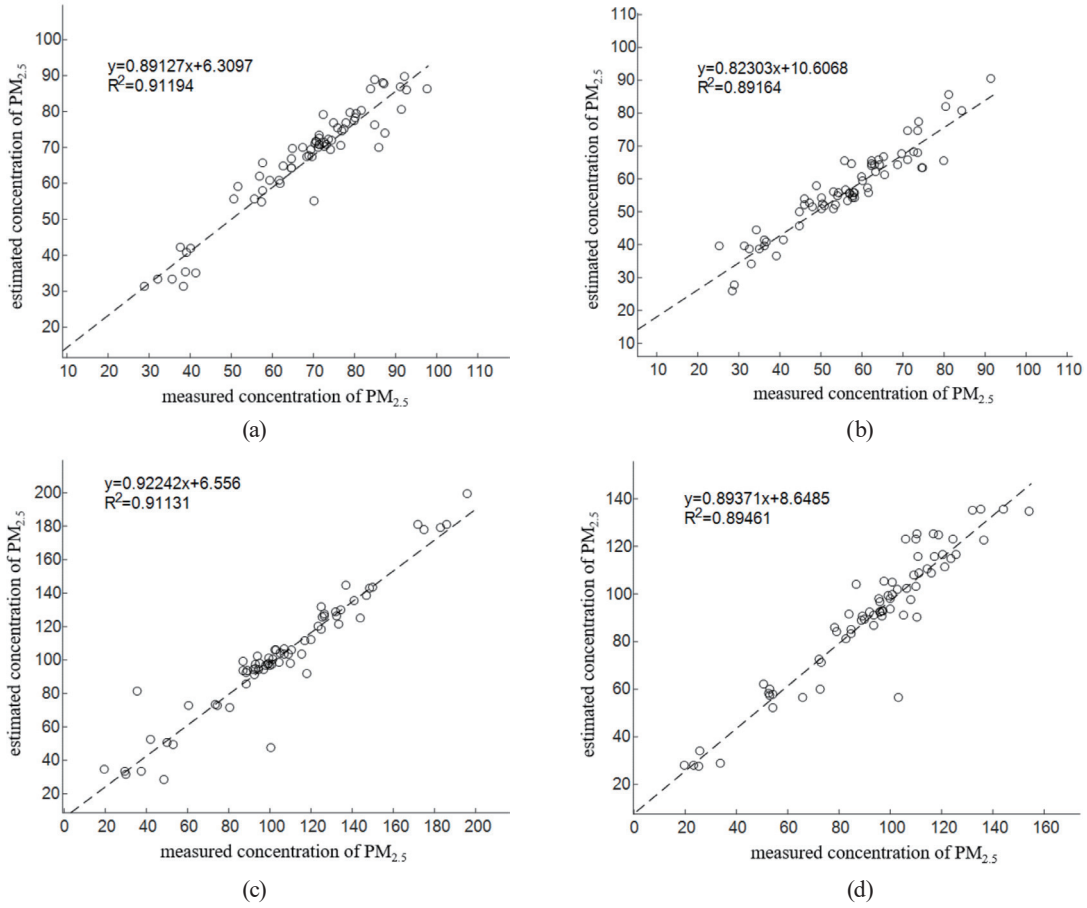

Fig. 4. Scatter diagrams showing correlation between ground-monitored $PM_{2.5}$ concentration and estimated concentration for S-GTWR: (a) spring, (b) summer, (c) autumn, and (d) winter.

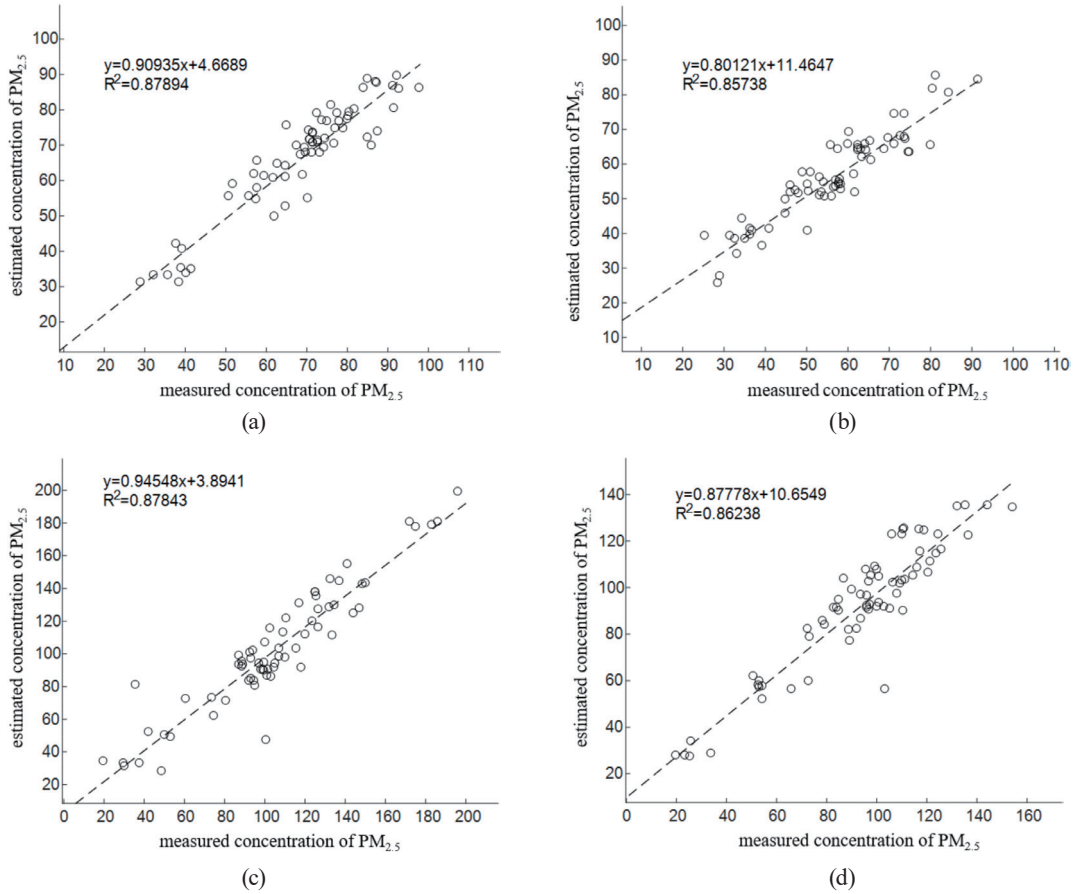

Fig. 5. Scatter diagrams showing correlation between ground-monitored and estimated $PM_{2.5}$ concentrations for GWR: (a) spring, (b) summer, (c) autumn, and (d) winter.

It was found with a 95% confidence level that the correlation between the estimated seasonal $PM_{2.5}$ concentration obtained by the S-GTWR model and the $PM_{2.5}$ concentration observed at ground monitoring stations was improved compared with that of GWR and kriging. The following conclusions were obtained by comparing the estimated $PM_{2.5}$ concentrations in different seasons obtained by the S-GTWR model. The R^2 values of all four models exceeded 0.89 [Figs. 4(a)–4(d)], while those of all four models of the other methods were below 0.88. The above results showed that the accuracy of the S-GTWR model for estimating the $PM_{2.5}$ concentration in each season is higher than those of the other models. The S-GTWR model can accurately estimate the concentration of $PM_{2.5}$ in the Beijing–Tianjin–Hebei region in different seasons from 2015 to 2019.

The difference between the measured and estimated $PM_{2.5}$ concentrations in each season and the cross-validation results of each model were compared. The conclusions were as follows. The most accurate model for estimating the $PM_{2.5}$ concentration in summer is the S-GTWR model; although the accuracy of the estimated $PM_{2.5}$ concentration of the winter model is relatively low, those of the spring and autumn models are high. Previous studies have shown that the contaminants from fossil fuel combustion, agricultural incineration, and automobile

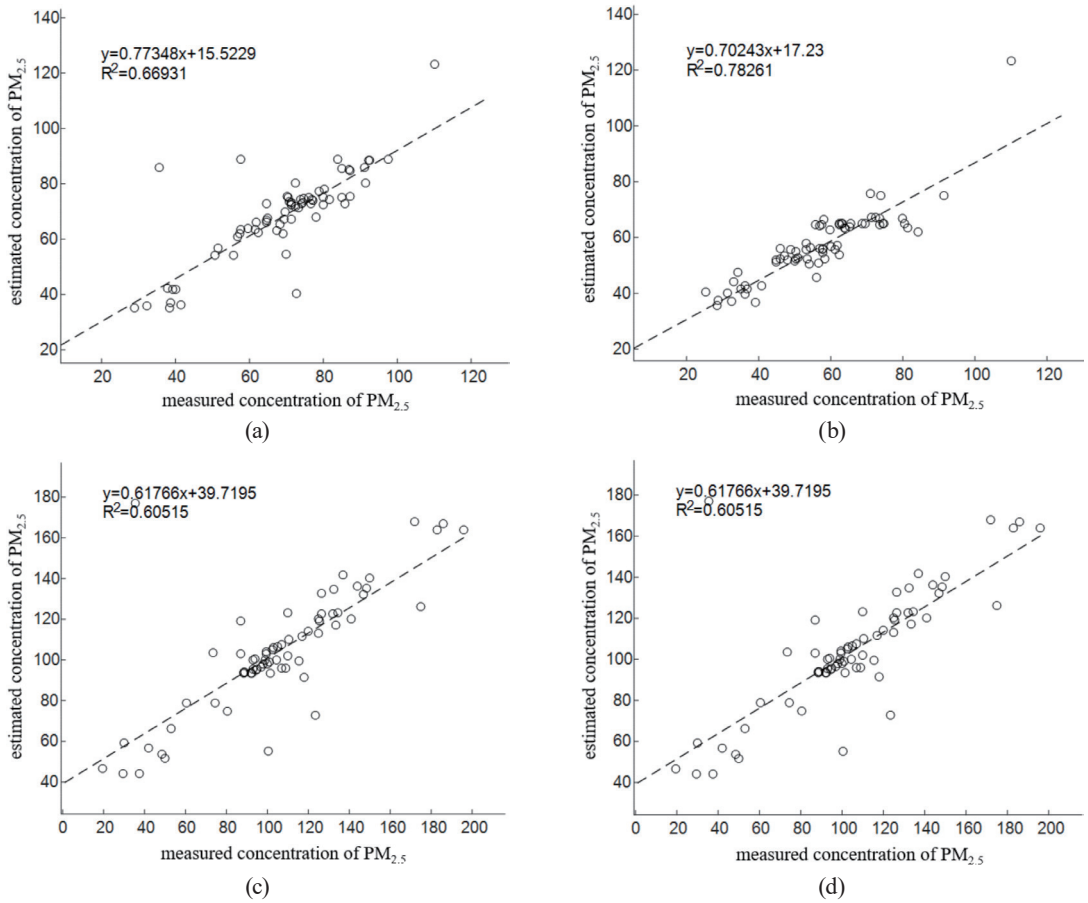

Fig. 6. Scatter diagrams showing correlation between ground-monitored and estimated $PM_{2.5}$ concentrations for kriging: (a) spring, (b) summer, (c) autumn, and (d) winter.

emissions account for more than 60% of the $PM_{2.5}$ concentration.[33] Autumn and winter are the periods of heating in northern cities. $PM_{2.5}$ pollutants released from fossil fuel combustion into the air increase nearly sevenfold during this period, directly leading to a significant increase in $PM_{2.5}$ concentration in autumn and winter.[34] At the same time, the different climatic characteristics in different seasons, such as differences in rainfall and temperature, have been shown to affect the production of $PM_{2.5}$ pollutants.[35,36] The Beijing–Tianjin–Hebei region has a temperate monsoon climate, with low temperatures, limited precipitation, high emissions, and the poor diffusion of pollutants in winter. These characteristics can cause the accumulation of $PM_{2.5}$ in the air. The concentration of $PM_{2.5}$ in winter is mainly affected by the climate, and natural factors such as meteorological factors are added to the evaluation model in this paper to study their effect on the accuracy of $PM_{2.5}$ concentration estimation. In summer, the high temperature and abundant rainfall are conducive to the volatilization and removal of $PM_{2.5}$. The probability of extreme weather problems in summer, such as sandstorms and dust, is very low, and there is no demand for heating. Therefore, the model established by using meteorological factors as characteristic variables to estimate the concentration of $PM_{2.5}$ has improved accuracy in summer.

Fine matter ($PM_{2.5}$) is a mixture of various chemical substances (such as water-soluble inorganic ions and organic matter) from natural and manmade sources.[37] A study of the chemical characterization and sources of $PM_{2.5}$ will help improve the $PM_{2.5}$ concentration estimation accuracy of the model. As the main components, chemical substances such as SO_2, NO_3, and NH_4 in fine particles play an essential role in the formation of fine particles.[38,39] Therefore, in the future, the sources and some chemical components of $PM_{2.5}$ in different seasons will be analyzed, and the analysis results will be added to the model for estimating the $PM_{2.5}$ concentration. To further evaluate the contribution of characteristic variables to the model in different seasons, in a follow-up study, the contribution of characteristic variables participating in the model in each season will be further measured, and the effect of each variable on the model accuracy will be analyzed.

3.2 Spatial validation

The characteristics of the spatiotemporal variation of the $PM_{2.5}$ concentration can be obtained from ground-level observations. However, these characteristics can only reflect the changes in $PM_{2.5}$ concentration near monitoring stations and cannot reflect the spatiotemporal distribution characteristics of $PM_{2.5}$ in the Beijing–Tianjin–Hebei urban agglomeration. Therefore, in this section, we discuss spatially continuous $PM_{2.5}$ concentration results obtained through simulation using different models, with which the characteristics of the continuous spatial distribution of the $PM_{2.5}$ concentration are studied. Figure 7 shows the spatially and

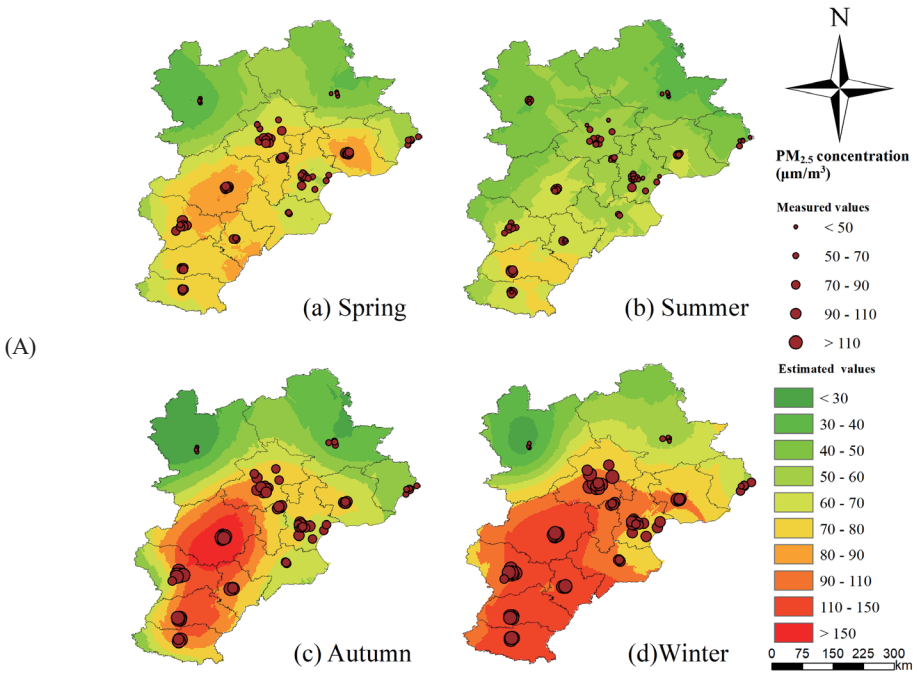

Fig. 7. (Color online) Spatially and temporally continuous distributions of seasonal mean of $PM_{2.5}$ concentrations obtained by GTWR and GWR models, and kriging for each season in Beijing–Tianjin–Hebei urban agglomeration from 2015 to 2019. (A) GTWR.

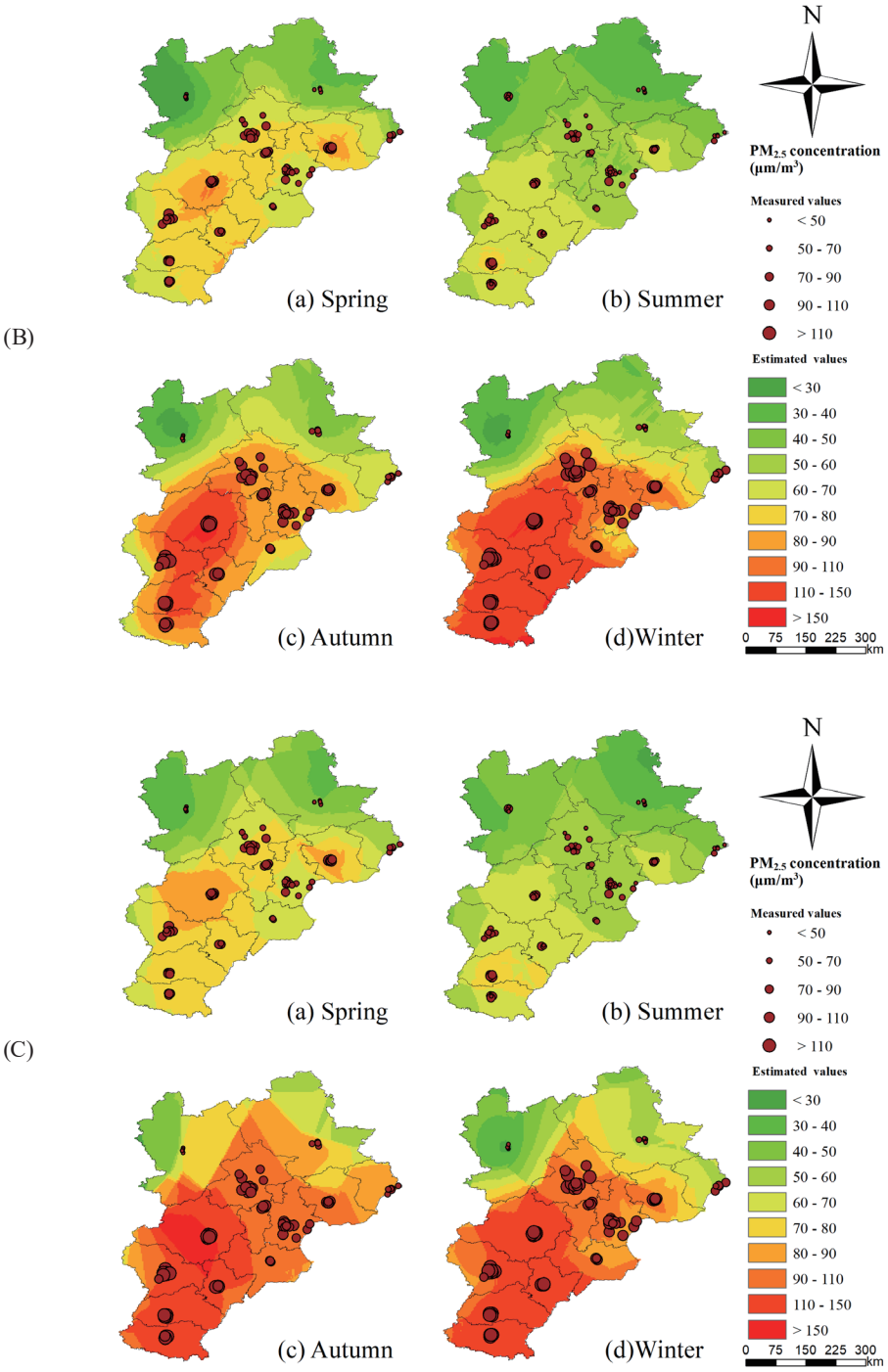

Fig. 7. (Color online) (Continued) (B) GWR and (C) kriging.

temporally continuous distributions of the seasonal mean $PM_{2.5}$ concentrations obtained by the GTWR and GWR models, and kriging for each season in the Beijing–Tianjin–Hebei urban agglomeration from 2015 to 2019.

Spatially, the $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei urban agglomeration generally gradually increases from north to south and from east to west. The $PM_{2.5}$ concentration in the northern part of the region is low, with the concentration in each season below 50 g/m³. The northern part of the Beijing–Tianjin–Hebei region is mostly mountainous with a high altitude. Under the comprehensive effects of terrain and topography, the airflow is relatively intense and most $PM_{2.5}$ pollutants move to the south. The coastal area in the east of the region has a low $PM_{2.5}$ concentration. This is because the amount of moisture in the air near the ocean increases at a higher air humidity, which increases the absorption of $PM_{2.5}$ pollutants. In addition, the coastal areas have strong wind waves and a high air mobility, which promote the rapid diffusion of $PM_{2.5}$ pollutants and play a role in alleviating $PM_{2.5}$ pollution. The southeastern part of the Beijing–Tianjin–Hebei region is mostly plain and the northern part of the region has a relatively high altitude, which make it difficult for $PM_{2.5}$ pollutants to disperse after gathering. The spatiotemporal distribution characteristics of the continuous $PM_{2.5}$ concentration obtained by the estimation models are consistent with those obtained from monitoring stations. The spatiotemporal distribution of the $PM_{2.5}$ concentration obtained by the GTWR model is more similar to the ground-measured observations than the distributions obtained by the GWR model and kriging. In addition, it can reflect details of the changes in $PM_{2.5}$ concentration and show relatively smooth changes. This is consistent with the actual $PM_{2.5}$ concentration distribution and its changes over the study area.

4. Conclusions

This study was conducted to accurately estimate the $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei region in different seasons based on data from both satellite sensors and ground-level observations. A multivariate variable group with optimal characteristics was selected by a greedy algorithm as the independent variables during model construction. The seasonal $PM_{2.5}$ concentration from 2015 to 2019 was estimated using the established GTWR model. Through the validation of the S-GTWR evaluation accuracy, the following conclusions were obtained: (1) the S-GTWR model with the optimal multivariate complex characteristic variables has good estimation performance, its RMSE is small, and R^2 exceeds 0.89 for every season. (2) By index comparison between the predicted and measured $PM_{2.5}$ concentrations, we found that the model has best performance for summer, followed by spring, autumn, and winter.

The climatic characteristics in winter have an impact on the $PM_{2.5}$ concentration distribution. Heating in winter produces a large amount of pollutant gases, which can strongly affect the $PM_{2.5}$ concentration distribution, thus reducing the accuracy of the model. Although the accuracy of the S-GTWR model is significantly improved compared with that of the existing model, there is still room for improvement in future research. For example, the accuracy and resolution of the input data will affect the performance of the model. The complex causes of the $PM_{2.5}$ concentration can be affected by natural and human factors, but the variables in the mathematical model and the availability of data for the $PM_{2.5}$ concentration estimation model are still limited. A more comprehensive and effective factor integration model will further improve the modeling accuracy and the understanding of the spatiotemporal distribution

characteristics of $PM_{2.5}$. Furthermore, the characteristic variables used for modeling also have uncertainty, which will affect the accuracy of the estimation results of the ground-level $PM_{2.5}$ concentration. All these issues will be addressed in future research.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant number 42077439), the National Key Research and Development Program of China (grant number 2018YFC0706003), and the Beijing Advanced Innovation Center for Future Urban Design (grant number UDC2018030611). The authors would like to thank the reviewers of the manuscript for their helpful comments. We are also grateful to the institutions named in this paper for providing remote sensing, environmental monitoring, and meteorological data.

References

1 Q. Zeng, J. Tao, L. Chen, H. Zhu, S. Zhu, and Y. Wang: Remote Sens. **12** (2020) 881. https://doi.org/10.3390/rs12050881
2 Y. Wang, M. Du, L. Zhou, G. Cai, and Y. Bai: Sustainability **11** (2019) 5713. https://doi.org/10.3390/su11205713
3 Y. Ju: J. Econ. Struct. **6** (2017). https://doi.org/10.1186/s40008-017-0089-4
4 G. Xu, X. Ren, K. Xiong, L. Li, X. Bi, and Q. Wu: Ecol. Indicators **110** (2020) 105889. https://doi.org/10.1016/j.ecolind.2019.105889
5 J. Guo, F. Xia, Y. Zhang, H. Liu, J. Li, M. Lou, J. He, Y. Yan, F. Wang, M. Min, and P. Zhai: Environ. Pollut. **221** (2017) 94. https://doi.org/10.1016/j.envpol.2016.11.043
6 Y. Xie, Y. Wang, K. Zhang, W. Dong, B. Lv, and Y. Bai: Environ. Sci. Technol. **49** (2015) 12280. https://doi.org/10.1021/acs.est.5b01413
7 C. H. M. Tong, S. H. L. Yim, D. Rothenberg, C. Wang, C.-Y. Lin, Y. D. Chen, and N. C. Lau: Atmos. Environ. **193** (2018) 79. https://doi.org/10.1016/j.atmosenv.2018.08.053
8 G. Geng, Q. Zhang, R. V. Martin, A. van Donkelaar, H. Huo, H. Che, J. Lin, and K. He: Remote Sens. Environ. **166** (2015) 262. https://doi.org/10.1016/j.rse.2015.05.016
9 M. D. Yazdi, Z. Kuang, K. Dimakopoulou, B. Barratt, E. Suel, H. Amini, A. Lyapustin, K. Katsouyanni, and J. Schwartz: Remote Sens. **12** (2020) 914. https://doi.org/10.3390/rs12060914
10 D. Voukantsis, K. Karatzas, J. Kukkonen, T. Rasanen, A. Karppinen, and M. Kolehmainen: Sci. Total Environ. **409** (2011) 1266. https://doi.org/10.1016/j.scitotenv.2010.12.039
11 H. Niska, M. Rantamäki, T. Hiltunen, A. Karppinen, J. Kukkonen, J. Ruuskanen, and M. Kolehmainen: Atmos. Environ. **39** (2005) 6524. https://doi.org/10.1016/j.atmosenv.2005.07.035
12 Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov: Atmos. Environ. **60** (2012) 632. https://doi.org/10.1016/j.atmosenv.2012.06.031
13 M. Menz, C. Gogu, S. Dubreuil, N. Bartoli, and J. Morio: Reliab. Eng. Syst. Saf. **196** (2020) 106771. https://doi.org/10.1016/j.ress.2019.106771
14 S. L. Zeger, D. Thomas, F. Dominici, J. M. Samet, J. Schwartz, D. Dockery, and A. Cohen: Environ. Health Perspect. **108** (2000). https://doi.org/10.1289/ehp.00108419
15 C. Brunsdon, A. S. Fotheringham, and M. E. Charlton: Geogr. Anal. **28** (1996). https://doi.org/10.1111/j.1538-4632.1996.tb00936.x
16 H.-J. Chu, B. Huang, and C.-Y. Lin: Atmos. Environ. **102** (2015) 176. https://doi.org/10.1016/j.atmosenv.2014.11.062
17 Y. Xue, Y. Li, J. Guang, A. Tugui, L. She, K. Qin, C. Fan, Y. Che, Y. Xie, Y. Wen, and Z. Wang: Remote Sens. **12** (2020) 855. https://doi.org/10.3390/rs12050855
18 B. Huang, B. Wu, and M. Barry: Int. J. Geogr. Inf. Sci. **24** (2010) 383. https://doi.org/10.1080/13658810802672469
19 J. M. Creamean, K. J. Suski, D. Rosenfeld, A. Cazorla, P. J. DeMott, R. C. Sullivan, A. B. White, F. M. Ralph, P. Minnis, J. M. Comstock, J. M. Tomlinson, and K. A. Prather: Science **339** (2013) 1572. https://doi.org/10.1126/science.1227279

20 M. Mirzaei, J. Amanollahi, and C. G. Tzanis: Air Qual. Atmos. Health **12** (2019) 1215. https://doi.org/10.1007/s11869-019-00739-z
21 Z. Fu and R. Li: Sci. Total Environ. **703** (2020) 135481. https://doi.org/10.1016/j.scitotenv.2019.135481
22 Q. Q. He and B. Huang: Remote Sens. Environ. **206** (2018) 72. https://doi.org/10.1016/j.rse.2017.12.018
23 Y. Guo, Q. Tang, D.-Y. Gong, and Z. Zhang: Remote Sens. Environ. **198** (2017) 140. https://doi.org/10.1016/j.rse.2017.06.001
24 M. Jiang, W. Sun, G. Yang, and D. Zhang: Remote Sens. **9** (2017) 346. https://doi.org/10.3390/rs9040346
25 N. Bei, X. Li, X. Tie, L. Zhao, J. Wu, X. Li, L. Liu, Z. Shen, and G. Li: Sci. Total Environ. **704** (2020) 135210. https://doi.org/10.1016/j.scitotenv.2019.135210
26 Y. Huang, Z. Li, H. Ye, S. Zhang, Z. Zhuo, A. Xing, and Y. Huang: Chin. Geogr. Sci. **29** (2019) 270. https://doi.org/10.1007/s11769-019-1027-1
27 Q. Wei, L. Zhang, W. Duan, and Z. Zhen: Int. J. Environ. Res. Public Health **16** (2019) 5107. https://doi.org/10.3390/ijerph16245107
28 H. J. Chu and M. Bilal: Environ. Sci. Pollut. Res. Int. **26** (2019) 1902. https://doi.org/10.1007/s11356-018-3763-7
29 M. Bakillah, R.-Y. Li, and S. H. L. Liang: Int. J. Geogr. Inf. Sci. **29** (2014) 258. https://doi.org/10.1080/13658816.2014.964247
30 A. Aziz, W. Osamy, A. M. Khedr, and A. Salim: J. King Saud Univ., Comp. Info. Sci. (2020) 1. https://doi.org/10.1016/j.jksuci.2020.03.010
31 Q. I. Xiaopeng, W. Liang, L. Barker, A. Lekiachvili, and Z. Xingyou: J. Resour. Ecol. **3** (2012) 220. https://doi.org/10.5814/j.issn.1674-764x.2012.03.004
32 Q. Yang, Z. Jiang, Z. Ma, and H. Li: Environ. Earth Sci. **72** (2014) 4303. https://doi.org/10.1007/s12665-014-3329-z
33 W. Guo, C. Long, Z. Zhang, N. Zheng, H. Xiao, and H. Xiao: Atmosphere **11** (2019) 5. https://doi.org/10.3390/atmos11010005
34 W. Guo, Z. Zhang, N. Zheng, L. Luo, H. Xiao, and H. Xiao: Atmos. Res. **234** (2020) 104687. https://doi.org/10.1016/j.atmosres.2019.104687
35 X. Huang, Z. Liu, J. Zhang, T. Wen, D. Ji, and Y. Wang: Atmos. Res. **168** (2016) 70. https://doi.org/10.1016/j.atmosres.2015.08.021
36 Q. He, Y. Yan, L. Guo, Y. Zhang, G. Zhang, and X. Wang: Atmos. Res. **184** (2017) 48. https://doi.org/10.1016/j.atmosres.2016.10.008
37 Y. Sun, G. Zhuang, Y. Wang, L. Han, J. Guo, M. Dan, W. Zhang, Z. Wang, and Z. Hao: Atmos. Environ. **38** (2004) 5991. https://doi.org/10.1016/j.atmosenv.2004.07.009
38 Y. Yang, R. Zhou, Y. Yu, Y. Yan, Y. Liu, Y. Di, D. Wu, and W. Zhang: J. Environ. Sci. (China) **55** (2017) 146. https://doi.org/10.1016/j.jes.2016.07.012
39 C. C. Meng, L. T. Wang, F. F. Zhang, Z. Wei, S. M. Ma, X. Ma, and J. Yang: Atmos. Res. **171** (2016) 133. https://doi.org/10.1016/j.atmosres.2015.12.013

About the Authors

Lei Zhou received his Ph.D. degree in cartography and geographical information systems from Academy of Disaster Reduction and Emergency Management, Beijing Normal University, Beijing, China, in 2011. He is currently an associate professor at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China. His research interests include spatiotemporal big data analysis, environmental remote sensing, earth surface processes, drought monitoring, and environmental quality comprehensive assessment.

Yani Wang received her M.D. in cartography and geographical information engineering from Beijing University of Civil Engineering and Architecture, Beijing, China. She is currently an assistant engineer with Gansu Institute of Natural Resources Planning and Research, Gansu. Her research interests include geographic information science and geospatial analysis.

Mingyi Du received his Ph.D. degree in geography from China University of Mining and Technology, Beijing, in 2001. He is currently working with Beijing University of Civil Engineering and Architecture (BUCEA). He is interested in refined urban operation management, Internet of Things technologies and applications, urban emergency management, and urban remote sensing.

Changfeng Jing received his B.S. degree in surveying and mapping from China University of Petroleum, in 2002, and his Ph.D. degree in geography from Zhejiang University, in 2008. Since 2008, he has been with Beijing University of Civil Engineering and Architecture (BUCEA), where he has been an associate professor at School of Geomatics and Urban Spatial Informatics since 2015. He is the author of two books and more than 30 articles, and has patented more than 10 inventions. His research interests include urban spatiotemporal analysis, geostatistics, urban Internet of Things, and urban planning management.

Siyu Wang is currently a graduate student at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China. Her research interests focus on earth surface processes and drought monitoring.

Congcong He is currently a graduate student at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China. Her research interests focus on earth surface processes based on remote sensing technology.

Ting Luo is currently a graduate student at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China. Her research interests focus on earth surface processes based on land use inversion using remote sensing.

Yinuo Zhu is currently an undergraduate student in geographical information systems at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China.

Ting Gao is currently an undergraduate student in geographical information systems at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China.

Kun Yang is currently a graduate student at School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China.