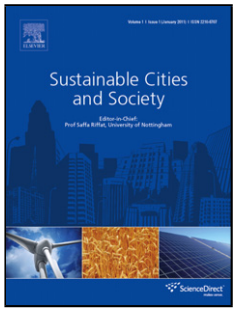
Journal Pre-proof



A Geographically Weighted Regression Model Augmented by Geodetector Analysis and Principal Component Analysis for the Spatial Distribution of $PM_{2.5}$

Rui Zhao, Liping Zhan, Mingxing Yao, Linchuan Yang

PII: S2210-6707(20)30093-7

DOI: https://doi.org/10.1016/j.scs.2020.102106

Reference: SCS 102106

To appear in: *Sustainable Cities and Society*

Received Date: 12 September 2019

Revised Date: 1 February 2020

Accepted Date: 17 February 2020

Please cite this article as: Zhao R, Zhan L, Yao M, Yang L, A Geographically Weighted Regression Model Augmented by Geodetector Analysis and Principal Component Analysis for the Spatial Distribution of $PM_{2.5}$, *Sustainable Cities and Society* (2020), doi: https://doi.org/10.1016/j.scs.2020.102106

A Geographically Weighted Regression Model Augmented by Geodetector Analysis and Principal Component Analysis for the Spatial Distribution of $PM_{2.5}$

Rui Zhao, Liping Zhan, Mingxing Yao, Linchuan Yang*

Southwest Jiaotong University, China

* Corresponding author at: Faculty of Architecture and Design, Southwest Jiaotong University, China.
E-mail address: yanglc0125@swjtu.edu.cn (L. Yang).

Highlights

- combine the use of Geodetector analysis, PCA, and the GWR model.
- present a novel approach to analyze complex spatially non-stationary relationships; (
- assess the spatial variation of $PM_{2.5}$ concentrations in the Pearl River Delta region by using the proposed methodology.

Abstract

This study develops an augmented geographically weighted regression (GWR) model to analyze the spatial distribution of $PM_{2.5}$ concentrations through the incorporation of Geodetector analysis and principal component analysis (PCA). The modeling approach we propose allows effective identification of important $PM_{2.5}$ drivers and their spatial variation. Technically speaking, Geodetector analysis is used to detect synergies between potential predictor and select predictor variables that truly affect the dependent variable, and PCA is adopted to eliminate multicollinearity among the variables. The spatial distribution of $PM_{2.5}$ concentrations within the Pearl River Delta region, China, is analyzed using the augmented GWR model. The augmented GWR model has an obvious advantage of parsimony, and moreover, it significantly outperforms the traditional regression model.

Keywords: Geographically weighted regression, Geodetector, Principal component analysis, $PM_{2.5}$,

collinearity, Pearl River Delta region, China

1. Introduction

In China, the urbanization-induced $PM_{2.5}$ pollution not only results in smog incidents but also poses a significant risk to human health. Therefore, it becomes a major environmental concern and its mitigation receives extensive attention from governments, scholars, and so forth (Li et al., 2019; Xu and Yang, 2019; Zhao et al., 2019). Monitoring $PM_{2.5}$ concentrations is the first and foremost step toward the identification of $PM_{2.5}$ sources and dynamics and can provide insights into $PM_{2.5}$ pollution mitigation and prevention (Pui et al., 2014). It is traditionally performed using ground (monitoring) stations. However, it is difficult to investigate spatial variations of $PM_{2.5}$ simply based on data from ground stations due to the limited number of stations and the non-standardized measurement frequency (Zou et al., 2015; Chu et al., 2016; Lee et al., 2016).

Numerous technological approaches, including geospatial interpolation (Li et al., 2014; Liu et al., 2014), the inversion of remote sensing data (Shi et al., 2018; Zhang et al., 2019), and geostatistical regression (Huang et al., 2018a), have been employed to fill the gaps by transforming discrete data points from ground stations into spatially distributed data. Importantly, spatial interpolations tend to be constrained by the number and geographical distribution of ground stations. If the distribution of ground stations is sparse and non-uniform, the accuracy and predictive power of the interpolation will decrease significantly (Li and Heap, 2011).

Remote sensing-based approaches usually utilize the aerosol optical depth (AOD) measured by a satellite to predict the concentration of near-surface atmospheric pollutants (Zhao et al., 2018). However, the ability to construct a relational model between the AOD and $PM_{2.5}$ is largely influenced by meteorological factors (Zheng et al., 2016). Geostatistical regression models, such as the geographically weighted regression (GWR) model, are often used to analyze the spatial variation or heterogeneity of atmospheric pollutant (e.g., $PM_{2.5}$ and NO_2) concentrations and its influencing factors (Hu et al., 2016; Wolf et al., 2017; Yang et al., 2017a). Notably, GWR models are exceptionally popular because they can account for spatially non-stationary (or spatially heterogeneous) relationships between the dependent variable and regressors (or predictor variables, independent variables) (Lloyd 2010; Harris et al. 2010a, 2011a; Harris and Juggins 2011; Xu and Huang, 2015; Wang et al., 2019; Yang et al., 2020).

An essential step of model construction is the identification and selection of variables to be included in the regression model. Existing methods that deal with variable selection are mainly based upon correlation analysis, supervised stepwise regression (Zhou et al., 2014), and cluster analysis (Yang et al., 2017b); however, these methods involve various degrees of uncertainty. For example, predictor variables may interact with each other, which may result in multicollinearity (Wu et al., 2015). However, if one cruelly excludes collinear variables, a loss of information may occur (Nguyen and Ng 2019; Tsao 2019).

To address the abovementioned issues, an augmented GWR model is developed in this study to describe the relationship between the spatial distribution of $PM_{2.5}$ concentrations and a set of contributory factors. Notably, Geodetector analysis is combined with principal component analysis (PCA) to determine key regressors for the GWR model. More specifically, Geodetector analysis reveals how variables interact with each other by analyzing their spatial disparities and is used to select important predictor variables (Wang et al., 2010, 2017); and PCA recombines the important predictor variables (selected by Geodetector analysis) into mutually independent regressors to reduce the multicollinearity (Zhai et al., 2018). The application of the two methods in tandem improves the representativeness of selected variables, thereby improving the predictive accuracy of the model and avoiding the multicollinearity problem. Following this approach, an augmented GWR model is developed to analyze the spatial distribution of $PM_{2.5}$ within the Pearl River Delta (*zhu jiang san jiao zhou*) region, China. Results illustrate the explanatory power of the augmented GWR model.

The contributions of this paper include the following: (1) combining Geodetector analysis, PCA, and the GWR model and presenting a novel approach to analyze complex spatially non-stationary relationships; (2) assessing the spatial variation of $PM_{2.5}$ concentrations in the Pearl River Delta region by using the proposed methodology.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on identifying the contributory factors of $PM_{2.5}$ concentrations and summarizes the factors. Section 3 introduces the study area and data. Section 4 reveals the methodologies of Geodetector analysis, PCA, and GWR modeling. Section 5 shows the results estimating from the augmented GWR model and presents technical discussions. Section 6 concludes the paper and points out future research directions.

2. Summary of factors affecting $PM_{2.5}$ concentrations

The identification of factors influencing $PM_{2.5}$ concentrations is critical for the construction of statistical models. The selection of influencing factors is often based upon the source apportionment of $PM_{2.5}$. $PM_{2.5}$ particles are emitted by a variety of sources, including industry, traffic, and domestic sources (Pui et al., 2014). Cheng et al. (2015) and Tian et al. (2015) noted that traffic emissions are the main source of $PM_{2.5}$ in Hong Kong and Chengdu, China, respectively and that they contribute to approximately 25% of the total in these regions. However, Hua et al. (2015) and Wang et al. (2016) found that industrial emissions are the main source of $PM_{2.5}$ in the Yangtze River Delta region and in Northwestern China, respectively and that they account for more than 30% of the total. In comparison, Yao et al. (2016) demonstrated that industrial emissions contribute up to 70% of the total $PM_{2.5}$ emissions in Northwestern China.

Existing findings indicate that the emission sources of $PM_{2.5}$ vary significantly across the region (Yang et al., 2018). In agricultural regions, the main source of $PM_{2.5}$ is the burning of straws and biological matters (Tao et al., 2013; Yin et al., 2017); in manufacturing and mining-dominated regions, the main source of $PM_{2.5}$ is industrial emissions (Wang et al., 2015); and in economically developed regions where the tertiary (service) sector is dominant, $PM_{2.5}$ emissions are closely related to domestic consumption and urban transport (Lin et al., 2014). Indeed, Ross et al. (2007) and Clougherty et al. (2008) found that the spatiotemporal variation in $PM_{2.5}$ concentrations is mainly caused by differences in land-use type and population distribution. Their conclusion is also supported by Wolf et al. (2017) and Lu et al. (2018).

Meteorological factors also play a major role in the creation of spatially varying $PM_{2.5}$ concentrations (Lin et al., 2015). Zhang et al. (2015) concluded that in Beijing, China, precipitation is negatively correlated with $PM_{2.5}$ concentrations, indicating that an increase in precipitation significantly decreases $PM_{2.5}$ concentrations. In the Sichuan Basin, southwestern China, Li et al. (2015) showed that temperature and atmospheric pressure affect the regional accumulation and migration of $PM_{2.5}$ by influencing the convective motion of the air. Chen et al. (2017) investigated the correlations between $PM_{2.5}$ concentrations and various meteorological factors in the Jing-Jin-Ji Metropolitan Region, Northern China. They found that wind speed has the greatest impact among all the contributory factors. Cheng et al. (2019) and Meng et al. (2019) showed that air temperature, atmospheric pressure, relative humidity, and precipitation are the most important factors that contribute to the spatial variation of $PM_{2.5}$ concentrations in Beijing.

By and large, these existing studies provide valuable insights into the identification of factors influencing the spatial distribution/variation of $PM_{2.5}$ concentrations. However, they do not adequately consider the potential influences of inter-factor interactions (Xu et al., 2014). We, therefore, propose an augmented GWR model to circumvent a few problems and limitations in existing methods. Moreover, following existing literature, independent variables used in this study are selected based on three key categories, namely meteorological conditions, urban geography factors, and $PM_{2.5}$ sources at two different levels (i.e., intragroup or intergroup). More specifically, land-use type and population counts are applied as the urban geographic factors, and industrial and traffic emissions are considered as the primary sources of $PM_{2.5}$.

3. Study area and data

3.1 Study area

The Pearl River Delta region (or the Pearl River Delta Metropolitan Region) is chosen as the study area. The region is located in southern Guangdong and also in the lower reaches of the Pearl River near the Southern China Sea, and it is the mainstay of in the Guangdong–Hong Kong–Macao Greater Bay Area (*yue gang ao da wan qu*). The region stretches from 112° E to 115.5° E and 21.5° N to 24° N and consists of 9 cities, namely Guangzhou, Huizhou, Dongguan, Shenzhen, Foshan, Zhongshan, Jiangmen, Zhaoqing, and Zhuhai (Fig. 1). The Pearl River Delta region is one of the most economically developed regions in China: its gross domestic product (GDP) accounts for 9.1% of the national total (National Bureau of Statistics, 2015).

The rapid economic development of the region inevitably increases stress on resources and the environment. The Pearl River Delta region is also one of the major air pollution control zones in China alongside the Jing-Jin-Ji (Beijing-Tianjin-Hebei) Metropolitan Region, the Yangtze River Delta Economic Zone, the Cheng-Yu (Chengdu-Chongqing) Economic Zone, and the Fen-Wei Plain.

3.2 Data sources

Data from various sources are used in this study, including ground-based $PM_{2.5}$ monitoring data, traffic data, industrial data, population density data, land-use data, and meteorological data (Table 1). The ground-based $PM_{2.5}$ monitoring data are obtained from the China National Environmental Monitoring Centre (NEMC, 2015) and recorded on an hourly basis from 1/1/2015 – 31/12/2015 by 54

air quality monitoring stations in 9 cities within the Pearl River Delta region. The hourly data are converted to annual average data for each station. The land-use data are derived from Landsat 8 remote sensing images in the Geospatial Data Cloud (GDC, 2015), and the following four types of land-use are considered: agricultural land, green space, building land, and water body. Monthly averages of each meteorological element recorded during 2015 at all the weather monitoring stations are obtained from the National Meteorological Information Center (NMIC, 2015). These data are subsequently processed to obtain annual average data. Traffic data are obtained from the BIGEM road vector dataset (http://www.bigemap.com/). Population density data (1 km × 1 km) are obtained from the Resource and Environment Data Cloud Platform (REDCP, 2015). Data related to industrial $PM_{2.5}$ emissions are obtained from field investigation, including the number of key enterprises that have been prioritized for air pollution monitoring within the study area and their emission loads.

It is noteworthy that the length of roads and the number of enterprises are used to assess traffic emissions and industrial emissions, respectively. Since data that directly reflect traffic intensity and industrial emissions are unavailable to the authors, the length of roads and the number of enterprises are used as proxies for such indicators. Moreover, previous studies have demonstrated the applicability of using the two factors to evaluate traffic emissions and industrial emissions because of their close relationships with $PM_{2.5}$. For instance, Hu et al. (2016) applied the length of roads in the study area as an indicator of traffic emissions and found that the indicator has a high correlation with $PM_{2.5}$. Similarly, Huang et al. (2017) screened predictors from a huge number of potential $PM_{2.5}$ influencing factors and used the length of roads as a measure of traffic emissions. Zhai et al. (2018) predicted the spatial distribution of $PM_{2.5}$ by incorporating the number of enterprises into the regression model.

3.3 Data integration

All of the datasets are projected onto the Gauss-Kruger/Beijing 1954 coordinate system in *ArcGIS (v 10.2)*.

Since the spatial position of the $PM_{2.5}$ monitoring stations does not perfectly match that of the meteorological stations, the meteorological data need to be processed to match the $PM_{2.5}$ data. However, interpolation involves the uncertainty problem. As such, several interpolation methods are

tested.

Relative errors of different interpolation methods for each meteorological measure are shown in Table 2. We opt for the method that provides the smallest relative error for each meteorological indicator in subsequent analysis: inverse distance weighting (IDW) interpolation is used for average and maximum wind speeds and average water vapor pressure; Kriging interpolation is used for relative humidity; natural neighbor interpolation is used for precipitation; and trend interpolation is used for average air temperature and average atmospheric pressure. Furthermore, several buffer zones (0.1 km – 10 km) are constructed around each $PM_{2.5}$ monitoring center (see Table 1), from which the datasets corresponding to each variable are extracted.

4. Methodology

4.1 Research framework

The augmented GWR model is constructed in several stages (Fig. 2). First, the five categories of factors in Table 1 are used as independent variables, while $PM_{2.5}$ concentrations are adopted as the dependent variable. Geodetector analysis (detailed in Section 4.2) is used to compute the contribution of each factor to $PM_{2.5}$ concentrations and to detect synergies between factors with respect to $PM_{2.5}$ concentrations. Since independent variables that put into Geodetector analysis should be categorical variables, it is necessary to categorize all continuous variables (Cao et al., 2013). Therefore, in this study, air temperature, atmospheric pressure, population density, and traffic sources are divided into 10 categories, while precipitation, water vapor pressure, and wind speed are divided into 9 categories. The correlation between each variable and $PM_{2.5}$ concentrations is assessed using Pearson's correlation analysis. Effective predictor variables are then selected based on the results of Pearson's correlation analysis and Geodetector analysis. Finally, PCA is used to select the principal components (PCs) whose cumulative contributions exceeded 95%. These PCs are used as input for the GWR model.

Model validation is performed using the leave-one-out cross-validation (LOOCV) method to orthogonally validate the model. For this, the datasets are divided into training and validation sets. The $PM_{2.5}$ concentration data and predictor variables at 53 monitoring stations in the training set are used to construct the GWR model, the results of which are used to predict the $PM_{2.5}$ concentration at the validation point. This process is repeated 54 times until every monitoring station has been used as the

validation point.

The augmented GWR model is compared to other regression models based on the coefficient of determination (R^2), the adjusted R^2, and the Akaike Information Criterion (AICc). Higher values of R^2 and adjusted R^2 and lower values of AICc are indicative of the improved model accuracy (Yang et al., 2019).

4.2 Geodetector analysis

The Geodetector method is a quantitative technique that determines whether the spatial distribution of a geostatistical variable is similar to that of an independent variable that has been identified as an important explanatory factor (Wang et al., 2010, 2017). The key idea behind Geodetector is that if factor *X* is associated with *Y*, then *X* and *Y* would exhibit similar spatial distributions. In other words, if the spatial variability of $PM_{2.5}$ concentrations is caused by a specific factor, there should be some similarity between the spatial distributions of the factor and $PM_{2.5}$ concentrations. The Geodetector method uses the power of determinant (q_X) to reflect the spatial correspondence of factor *X* and *Y* by using the following equation (Wang et al., 2010, 2017):

$$q_X = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}, \quad (1)$$

where *N* is the number of samples in the study area; N_h is the number of samples in zone (category) *h* of factor *X*; σ^2 is the total variance of *Y* in the study area; σ_h^2 is the variance of *Y* within zone (category) *h* of factor *X*; and *L* is the number of zones (categories) of factor *X*. $N_h \sigma_h^2$ is within sum of variances, and $N\sigma^2$ is total sum of variances. The greater the value of q_X is, the more factor *X* explains *Y*, and *vice versa*.

If $q(X_1 \cap X_2) > \mathrm{Max}(q(X_1), q(X_2))$, then the interactions between X_1 and X_2 increase their influences on *Y*. Conversely, if $q(X_1 \cap X_2) < \mathrm{Max}(q(X_1), q(X_2))$, then the influence of X_1 and X_2 on *Y* is diminished by the interactions between X_1 and X_2. Finally, if $q(X_1 \cap X_2) = q(X_1) + q(X_2)$, then the effects of X_1 and X_2 on *Y* are mutually independent (Wang et al., 2010, 2017).

4.3 Principal component analysis

PCA uses an orthogonal transformation to convert possibly correlated variables into a number of linearly uncorrelated variables, namely independent PCs (Abdul-Wahab et al., 2005; Harris et al. 2011b; Demšar et al., 2013). The first PC has the largest possible variance, constituting as much of the variability in the data as possible, following by the second PC and the third PC. In this way, PCA reduces collinearity between predictor variables. Each PC can be expressed as follows:

$$PC_i = l_{i1}X_1 + l_{i2}X_2 + \cdots + l_{in}X_n, \quad (2)$$

where PC_i is the i-th PC; X_j is the j-th predictor variable; and l_{ij} is the coefficient of X_j ($i, j = 1, 2 \ldots, n$).

4.4 GWR model

Traditional regression models use only one equation to describe the global relationships between predictor and predicted variables and produce one-size-fits-all outcomes. However, they ignore some interesting and important local differences in the determinants (Fotheringham et al., 2002). By contrast, GWR, a local regression technique, relaxes the assumption of constant (or spatially invariant) relationships between predictor and predicted variables in traditional regression models and creates multiple equations to describe such relationships. GWR models can reflect relationships with a space-varying nature and offer local or location-specific regression results (Harris et al., 2014).

GWR is seminally proposed by Brunsdon et al. (1996) and Fotheringham et al. (2002). Since then, it has become a popular approach to modeling social processes. GWR is a locally linear regression technique that captures spatially non-stationary relationships between the dependent variable and predictor variables by incorporating geographical information. In other words, the GWR model predicts the relationship between the dependent variable and predictor variables at each location by constructing local (or location-specific) regression equations. The GWR model is mathematically expressed as follows (Fotheringham et al., 2002):

$$Y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) X_{ik} + \varepsilon_i, \quad (3)$$

where Y_i is the dependent variable; X_{ik} is the k-th independent variable; (u_i, v_i) is the geographical coordinates of the i-th point; $\beta_0 (u_i, v_i)$ is the intercept of the i-th point; $\beta_k (u_i, v_i)$ is the coefficient of X_{ik} and ε_i is the residual of the i-th point.

A regression equation that considers only nearby observations is developed for each point by using a weighted least square method. Each nearby observation is weighted by means of a function of

the distance to the regression point through various methods. Common spatial weighting or distance decay methods include fixed Gaussian and adaptive bi-square kernel functions. The fixed Gaussian function can be written as

$$w_{ij} = \exp(-d_{ij}^{\ 2} / \theta^2), \tag{4}$$

where w_{ij} is the weight value of observation *j* for estimating the coefficient of observation *i*, d_{ij} is the straight-line distance between observations *i* and *j*, and θ is a constant bandwidth.

The adaptive bi-square kernel function allows the spatial extent to vary (or keeps adjusting the spatial extent) at different regression points and includes the same number of neighbors for local model estimation. It can be written as

$$w_{ij} = \begin{cases} (1 - d_{ij}^{\ 2} / \theta_{i(k)}^{\ 2})^2 & d_{ij} < \theta_{i(k)} \\ 0 & d_{ij} > \theta_{i(k)} \end{cases}, \tag{5}$$

where $\theta_{i(k)}$ is an adaptive bandwidth; and other variables are as defined before.

Interest readers can refer to Brunsdon et al. (1996) and Fotheringham et al. (2002) for more information of GWR.

5. Results and Discussion

5.1 Results from the augmented GWR model

Table 3 shows the correlations between variables from Geodetector analysis. It is clear that the meteorological factors are important determinants of $PM_{2.5}$ concentrations. In addition, inter-factor interactions significantly increase $PM_{2.5}$ levels. By performing Pearson's correlation analysis, 14 effective predictor variables are selected from 119 potential factors, as shown in Table 4. However, the high value of the variance inflation factor (VIF) indicates a significant degree of multicollinearity between these variables.

To eliminate multicollinearity, PCA is used to convert the 14 variables into 14 mutually independent PCs (PC_1 – PC_{14}), through which the VIF values of predictor variables are significantly reduced. Table 5 reveals PCA results. PC_1, PC_2, ... , and PC_8 carry over 95% of the total variance, and they are used in subsequent analysis. Other PCs are discarded.

Both fixed Gaussian and adaptive bi-square kernel functions for GWR modeling were tested in this study, and they produce very similar results. As such, finally, the adaptive bi-square kernel function was used as the distance decay function, and the Golden bandwidth selection method was adopted in the GWR estimation. The fit and cross-validation results of the proposed GWR model are shown in Table 6. The R^2 values indicate that the goodness-of-fit of the model is excellent, indicating that the predictor variables explaining 84% of the variance in $PM_{2.5}$ concentrations. In addition, the mean absolute prediction error (MAE), mean relative error (MRE), and root mean square error (RMSE) of the model are all relatively small. Thus, we can conclude that the augmented GWR model is stable and robust.

Comparing the augmented GWR model and other models reported in existing literature can more or less provide insights into model performance. Ding et al. (2016) employed kriging interpolation, IDW interpolation, and regularized spline interpolation to interpolate $PM_{2.5}$ concentrations for the Pearl River Delta region. These techniques achieved MREs of 0.195, 0.192, and 0.186, and RMSEs of 10.781, 10.295, and 10.142, respectively. The values of MREs and RMSEs are higher than those of the augmented GWR model (see Table 6). This can be explained by the fact that the augmented GWR model accounts for spatially non-stationarity relationships of predictor variables, overcoming an obvious inherent weakness of non-spatial (or aspatial) regression, namely the omission of geographic information.

Fig. 3 shows the spatial distribution of $PM_{2.5}$ concentrations in the Pearl River Delta region derived from the augmented GWR model. The maximum annual average $PM_{2.5}$ concentration for the Pearl River Delta region is 39.05 $\mu g/m^3$, which is far lower than that of the Jing-Jin-Ji Metropolitan Region (125.54 $\mu g/m^3$) (Zhai et al., 2018). This can be attributed to the Air Pollution Prevention and Control Action that was promulgated in 2013 by the Central Government of China. This plan calls for a 15% reduction in $PM_{2.5}$ levels by 2017 relative to 2012 levels, the use of energy-saving measures, and the upgrading in all coal-fired power plants, coal-fired boilers, and industrial furnaces within this region by the end of 2015 (China's State Council, 2013). The enactment of this plan leads to the optimization of energy structures and industrial transformation throughout the Pearl River Delta, which saw a 17% reduction in $PM_{2.5}$ concentrations from 46 $\mu g/m^3$ to 38 $\mu g/m^3$ by the end of 2015 (Jiang et al., 2015).

Annual average $PM_{2.5}$ concentrations are generally higher in the northwest of the region and lower in the southeast. $PM_{2.5}$ concentrations are significantly lower in the southeast (e.g., Shenzhen and Dongguan) than in the central region (e.g., Foshan and Guangzhou). This is consistent with the findings of Shen and Yao (2017). Possible explanations for this spatial pattern are as follows: (1) Shenzhen and Huizhou are coastal cities with highly variable land-cover types. This feature encourages the formation of small, local circulations between the sea and land, thereby increasing the atmosphere's diffusion capacity and decreasing the $PM_{2.5}$ concentrations (Yan et al., 2018); (2) As Guangzhou and Foshan are far from the shoreline, the atmospheric dispersion is relatively limited in these regions. Furthermore, these industrially developed and densely populated cities account for a large share of the Pearl River Delta's total $PM_{2.5}$ emissions (Huang et al., 2018b); (3) Between the nine major cities in the Pearl River Delta region, Shenzhen and Zhuhai have higher levels of "green development" due to the transformation of their economic structures and the vigorous development of high-tech industries (Wang et al., 2020). In contrast, Zhaoqing and Jiangmen have relied on traditional manufacturing and processing for a long time, and the growth of emerging industries in these cities is relatively limited resulting in significant environmental pressures. This implies that the contributions of the industrial and energy sectors to $PM_{2.5}$ concentrations vary significantly due to differences in the economic structure (Wang et al., 2018).

5.2 Model comparison

The performance of the augmented GWR model is compared to an ordinary least squares (OLS) regression model and a conventional GWR model using the same datasets. Results are shown in Table 7. The two GWR models outperformed the OLS model in all three measures.

The augmented GWR model has an adequate goodness-of-fit that is close to that of the conventional GWR model. However, in this study, the conventional GWR model uses numerous variables and suffers from the multicollinearity problem (shown by high VIF values). In other words, simply using the conventional GWR model is not rigorous, and the result estimating from the conventional GWR model is not reliable. Fortunately, the augmented GWR model does not have such a problem and uses much fewer variables. It has a distinct strength of parsimony, and its performance is good.

The augmented GWR model for the spatial distribution of the Pearl River Delta region has a high goodness-of-fit relative to its counterparts reported in existing studies. For example, Song et al. (2014) used a general linear regression model, a semi-empirical model, and a GWR model to simulate the spatial distribution of $PM_{2.5}$ concentrations in the Pearl River Delta region. The adjusted R^2 values for the three models are 0.564, 0.526 and 0.74, respectively, which are lower than for our augmented GWR model developed in this study. Furthermore, Yang et al. (2017c) utilized ground monitoring data, satellite remote sensing, air quality model, and geographic and local source related spatial inputs to predict the distribution of $PM_{2.5}$ concentrations in the Pearl River Delta region. They reported an adjusted R^2 of 0.676, which is also smaller than that of our augmented GWR model. Admittedly, the difference in the model performance in previous studies can be related to the selection of predictor factors. Song et al. (2014) mainly focused on the impacts of meteorological factors, while Yang et al. (2017c) placed an emphasis on source apportionment.

5.3 Discussion

From our case study, the augmented GWR model showed a high goodness-of-fit, considerably larger than that of the conventional OLS model; moreover, it performs similarly to the convention GWR model that utilizes much more variables (than the augmented GWR model), suffers from the multicollinearity problem, and thus produces unreliable results. The augmented GWR model has an obvious strength of parsimony. All the above indicates possible advantages of the augmented GWR model in modeling the spatial distribution of air pollutants on a regional scale, even when ground monitoring stations are unevenly distributed.

Although the augmented GWR model is applicable for modeling the spatial distribution of $PM_{2.5}$, there are a number of uncertainties (Propastin et al., 2008) remained: (1) Uncertainties in the data source of $PM_{2.5}$: First, the distribution of $PM_{2.5}$ ground stations are uneven: the number of stations in the middle area is more than that in the surrounding area (see Fig. 1). Moreover, the meteorological data has been processed to maintain on the same measurement scale compared with the $PM_{2.5}$ concentrations. Though several interpolation methods have been tested, uncertainties still occurred at their data integration; (2) Uncertainties in the augmented GWR model itself: This study models spatially non-stationary relationships between $PM_{2.5}$ concentrations and a number of potential contributory factors. However, the formation of $PM_{2.5}$ and its composition and physicochemical

characteristics are omitted, which would have an unknown impact on the explanatory power of the model; and (3) Uncertainties in PCA: To avoid the multicollinearity problem, variables used in the augmented GWR model are selected by using PCA, in which some of the information losses. This gives rise to uncertainties in model results.

When assessing GWR as a predictor, Kriging with an external drift (KED) (alternatively called universal kriging) offers a useful alternative to GWR from the Geostatistics paradigm. KED caters for solving the spatially non-stationarity among variables, which considers additional information as external drift (Harris et al., 2010b, 2011a). However, KED usually selects variables that show high correlations with PM2.5 as the external drift (Pearce et al., 2009; Ramos et al., 2016). Therefore, it is uncertain whether KED has better performance on the prediction by using the Geodetector and PCA to select key variables among a huge number of possible influencing factors. In addition, as mentioned above, the distribution of $PM_{2.5}$ ground monitoring stations are uneven in the study area: the number of stations in the middle area is more than that in the surrounding area. For this reason, this study prefers to use geostatistical regression instead of spatial interpolation to improve the coverage from individual points to broader planes. However, testing the power of KED for the studied problem is left for future research.

6. Conclusions

In this study, a GWR model is augmented by the incorporation of Geodetector analysis and PCA to better characterize the spatial distribution of $PM_{2.5}$ concentrations in the Pearl River Delta region, China. Geodetector analysis is used to assess the contribution of potential influencing factors and their interactions with $PM_{2.5}$ concentrations, and PCA is used to eliminate multicollinearity between the factors. The augmented GWR model is capable of identifying the contribution of each factor, thus ensuring that selected variables produced a model with a great predictive capacity than other modeling approaches; the augmented GWR model achieved reasonable goodness-of-fit, much higher than the conventional OLS model. Our model simulation results indicate that there are significant disparities between the eastern and western regions of the Pearl River Delta region with respect to $PM_{2.5}$ concentrations in 2015; and that $PM_{2.5}$ concentrations are relatively high in western cities such as Zhaoqing, Foshan, Guangzhou, and Jiangmen, but relatively low in the southeast of the region, which includes Shenzhen, Dongguan, Huizhou, Zhuhai, and Zhongshan.

Future work is needed to further validate the proposed GWR model, including its suitability for different spatial and temporal resolutions, convergence in long-term predictions, and sensitivity towards changes in the regional scale. In addition, variables used in the model can be updated based on an improved understanding and data on $PM_{2.5}$ formation mechanisms, as well as its composition and physicochemical characteristics, to improve predictive accuracy. Furthermore, as noted above, KED, the power of which has extensively confirmed in existing literature (Harris et al. 2010b, 2011a; Pearce et al., 2009; Ramos et al., 2016), should be explored for the studied problem in upcoming research.

Acknowledgements

This study is sponsored by the National Natural Science Foundation of China (No.41571520), Sichuan Provincial Key Technology Support (No.2017SZ0169; No. 2019JDJQ0020), and Sichuan Province Circular Economy Research Center Fund (No. XHJJ-1802). The authors are grateful to the three reviewers for their constructive comments.

References

Abdul-Wahab, S.A., Bakheit, C.S., & Al-Alawi, S.M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10), 1263-1271.

Brunsdon, C., Fotheringham, A.S., & Charlton, M.E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281-298.

Cao, F., Ge, Y., & Wang, J.F. (2013). Optimal discretization for geographical detectors-based risk assessment. *GIScience Remote Sensing*, 50(1), 78-92.

Chen, Z., Cai, J., Gao, B., Xu, B., Dai, S., He, B., & Xie, X. (2017). Detecting the causality influence of individual meteorological factors on local $PM_{2.5}$ concentration in the Jing-Jin-Ji region. *Scientific Reports*, 7, 40735.

Cheng, Y., Lee, S., Gu, Z., Ho, K., Zhang, Y., Huang, Y., Chow, J.C., Watson, J.G., Cao, J., & Zhang, R. (2015). $PM_{2.5}$ and $PM_{10\text{-}2.5}$ chemical composition and source apportionment near a Hong Kong roadway. *Particuology*, 18, 96-104.

Cheng, N., Cheng, B., Li, S., & Ning, T. (2019). Effects of meteorology and emission reduction measures on air pollution in Beijing during heating seasons. *Atmospheric Pollution Research*, 10(3), 971-979.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., & Xiang, H. (2016). A review on predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth. *Atmosphere*, 7(10), 129.

Clougherty, J.E., Wright, R.J., Baxter, L.K., & Levy, J.I. (2008). Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants. *Environmental Health*, 7(1), 17.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, *103*(1), 106-128.

Ding, H., Yu, Z., Xu, W., Cao, S., Li, H., & Liu, Y. (2016). Comparative study of the three spatial interpolation methods for the regional air quality evaluation. *Journal of Safety and Environment*, 91, 124-132. (In Chinese with English Abstract).

Fotheringham, A.S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The*

Analysis of Spatially Varying Relationships. Chichester: Wiley.

GDC, 2015. Geospatial Data Cloud. Available online: http://www.gscloud.cn/.

Harris, P., Fotheringham, A. S., & Juggins, S. (2010a). Robust geographically weighted regression: a technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes. *Annals of the Association of American Geographers*, *100*(2), 286-306.

Harris, P., Fotheringham, A. S., Crespo, R., & Charlton, M. (2010b). The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. *Mathematical Geosciences*, *42*(6), 657-680.

Harris, P., Brunsdon, C., & Fotheringham, A. S. (2011a). Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. *Stochastic environmental Research and Risk Assessment*, *25*(2), 123-138.

Harris, P., Brunsdon, C., & Charlton, M. (2011b). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717-1736.

Harris, P., & Juggins, S. (2011). Estimating freshwater acidification critical load exceedance data for Great Britain using space-varying relationship models. *Mathematical Geosciences*, *43*(3), 265.

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., & Clarke, A. (2014). Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences*, *46*(1), 1-31.

Hu, L., Liu, J., & He, Z. (2016). Self-adaptive revised land use regression models for estimating $PM_{2.5}$ concentrations in Beijing, China. *Sustainability*, 8(8), 786.

Hua, Y., Cheng, Z., Wang, S., Jiang, J., Chen, D., Cai, S., Fu, X., Fu, Q., Chen, C., Xu, B., & Yu, J. (2015). Characteristics and source apportionment of $PM_{2.5}$ during a fall heavy haze episode in the Yangtze River Delta of China. *Atmospheric Environment*, 123, 380-391.

Huang, Y., Yan, Q., & Zhang, C. (2018a). Spatial–temporal distribution characteristics of $PM_{2.5}$ in China in 2016. *Journal of Geovisualization and Spatial Analysis*, 2(2), 12.

Huang, Y., Deng, T., Li, Z., Wang, N., Yin, C., Wang, S., & Fan, S. (2018b). Numerical simulations for the sources apportionment and control strategies of $PM_{2.5}$ over Pearl River Delta, China, part I: inventory and $PM_{2.5}$ sources apportionment. *Science of the Total Environment*, 634, 1631-1644.

Jiang, X., Hong, C., Zheng, Y., Zheng, B., Guan, D., Gouldson, A., Zhang, Q., & He, K. (2015). To what extent can China's near-term air pollution control policy protect air quality and human health? A case study of the Pearl River Delta region. *Environmental Research Letters*, 10, 104006.

Lee, M., Kloog, I., Chudnovsky, A., Lyapustin, A., Wang, Y., Melly, S., Coull, B., Koutrakis, P., & Schwartz, J. (2016). Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011. *Journal of Exposure Science and Environmental Epidemiology*, 26(4), 377-384.

Li, J., & Heap, A.D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6(3-4), 228-241.

Li, L., Losser, T., Yorke, C., & Piltner, R. (2014). Fast inverse distance weighting-basedspatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the contiguous us using parallel programming and kd tree. *International Journal of Environmental Research and Public Health*, 11(9), 9101-9141.

Li, Y., Chen, Q., Zhao, H., Wang, L., & Tao, R. (2015). Variations in PM_{10}, $PM_{2.5}$ and $PM_{1.0}$ in an urban area of the Sichuan Basin and their relation to meteorological factors. *Atmosphere*, 6(1), 150-163.

Li, J., Liao, H., Hu, J., & Li, N. (2019). Severe particulate pollution days in China during 2013–2018 and the associated typical weather patterns in Beijing-Tianjin-Hebei and the Yangtze River Delta regions. *Environmental Pollution*, 248, 74-81.

Lin, G., Fu, J., Jiang, D., Hu, W., Dong, D., Huang, Y., & Zhao, M. (2014). Spatio-temporal variation of $PM_{2.5}$ concentrations and their relationship with geographic and socioeconomic factors in China. *International Journal of Environmental Research and Public Health*, 11(1), 173-186.

Lin, G., Fu, J., Jiang, D., Wang, J., Wang, Q., & Dong, D. (2015). Spatial variation of the relationship between $PM_{2.5}$ concentrations and meteorological parameters in China. *Biomed Research International*, 2015, 259-265.

Liu, R., Chen, Y., Sun, C., Zhang, P., Wang, J., Yu, W., & Shen, Z. (2014). Uncertainty analysis of total phosphorus spatial–temporal variations in the Yangtze River Estuary using different interpolation methods. *Marine Pollution Bulletin*, 86(1-2), 68-75.

Lloyd, C. D. (2010). Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *30*(3), 390-405.

Lu, D., Mao, W., Yang, D., Zhao, J., & Xu, J. (2018). Effects of land use and landscape pattern on $PM_{2.5}$ in Yangtze River Delta, China. *Atmospheric Pollution Research*, 9(4), 705-713.

Meng, C., Cheng, T., Gu, X., Shi, S., Wang, W., Wu, Y., & Bao, F. (2019). Contribution of meteorological factors to particulate pollution during winters in Beijing. *Science of the Total Environment*, 656, 977-985.

National Bureau of Statistics, 2015. National Bureau of Statistics. Available online: http://www.stats.gov.cn/.

NEMC, 2015. China National Environmental Monitoring Centre. Available online: http://www.cnemc.cn/.

Nguyen, V.C., & Ng, C.T. (2019). Variable selection under multicollinearity using modified log penalty. *Journal of Applied Statistics*, doi: 10.1080/02664763.2019.1637829

NMIC (National Meteorological Information Center), 2015. Available online: http://data.cma.cn/.

Pearce, J.L., Rathbun, S.L., Aguilar-Villalobos, M., & Naeher, L.P. (2009). Characterizing the spatiotemporal variability of $PM_{2.5}$ in Cusco, Peru using kriging with external drift. Atmospheric Environment, 43(12), 2060-2069.

Pui, D.Y., Chen, S.C., & Zuo, Z. (2014). $PM_{2.5}$ in China: measurements, sources, visibility and health effects, and mitigation. *Particuology*, 13, 1-26.

Ramos, Y., St-Onge, B., Blanchet, J.P., & Smargiassi, A. (2016). Spatio-temporal models to estimate daily concentrations of fine particulate matter in Montreal: Kriging with external drift and inverse distance-weighted approaches. Journal of Exposure Science & Environmental Epidemiology, 26(4), 405-414.

REDCP, 2015. Resource and Environment Data Cloud Platform. Available online: http://www.resdc.cn/.

Ross, Z., Jerrett, M., Ito, K., Tempalski, B., & Thurston, G.D. (2007). A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environment*, 41(11), 2255-2269.

Shen, Y., & Yao, L. (2017). $PM_{2.5}$, population exposure and economic effects in urban agglomerations of China using ground-based monitoring data. *International Journal of Environmental Research and Public Health*, 14(7), 716.

Shi, Y., Ho, H.C., Xu, Y., & Ng, E. (2018). Improving satellite aerosol optical Depth-$PM_{2.5}$ correlations using land use regression with microscale geographic predictors in a high-density urban context. *Atmospheric Environment*, 190, 23-34.

Song, W., Jia, H., Huang, J., & Zhang, Y. (2014). A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the Pearl River Delta region in China. *Remote Sensing of Environment*, 154, 1-7.

Tao, J., Zhang, L., Engling, G., Zhang, R., Yang, Y., Cao, J., Zhu, C., Wang, Q., & Luo, L. (2013). Chemical composition of $PM_{2.5}$ in an urban environment in Chengdu, China: importance of springtime dust storms and biomass burning. *Atmospheric Research*, 122, 270-283.

Tian, Y.Z., Shi, G.L., Han, B., Wu, J.H., Zhou, X.Y., Zhou, L.D., Zhang, P., & Feng, Y.C. (2015). Using an improved source directional apportionment method to quantify the $PM_{2.5}$ source contributions from various directions in a megacity in China. *Chemosphere*, 119, 750-756.

Tsao, M. (2019). Estimable group effects for strongly correlated variables in linear models. *Journal of Statistical Planning and Inference*, 198, 29-42.

Wang, J.F., Li, X.H., Christakos, G., Liao, Y.L., Zhang, T., Gu, X., & Zheng, X.Y. (2010). Geographical detectors- based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24(1), 107-127.

Wang, J.F., & Xu, C.D. (2017). Geodetector: principle and prospective. *Acta Geographica Sinica*, 72(1), 116-134.

Wang, L., Wei, Z., Wei, W., Fu, J.S., Meng, C., & Ma, S. (2015). Source apportionment of $PM_{2.5}$ in top polluted cities in Hebei, China using the CMAQ model. *Atmospheric Environment*, 122, 723-736.

Wang, Y., Jia, C., Tao, J., Zhang, L., Liang, X., Ma, J., Gao, H., Huang, T., & Zhang, K. (2016). Chemical characterization and source apportionment of $PM_{2.5}$ in a semi-arid and petrochemical-industrialized city, Northwest China. *Science of the Total Environment*, 573, 1031-1040.

Wang, M.X., Zhao, H.H., Cui, J.X., Fan, D., Lv, B., Wang, G., Li, Z.H., & Zhou, G.J. (2018). Evaluating green development level of nine cities within the Pearl River Delta, China. *Journal of Cleaner Production*, 174, 315-323.

Wang, J., Wang, S., & Li, S. (2019). Examining the spatially varying effects of factors on $PM_{2.5}$ concentrations in Chinese cities using geographically weighted regression modeling. *Environmental Pollution*, 248, 792-803.

Wang, X., Chan, C. K. C., & Yang, L. (2020). Economic upgrading, social upgrading, and Chinese rural migrant workers in the Pearl River Delta. *China Review: An Interdisciplinary Journal on*

Greater China, in press.

Wolf, K., Cyrys, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., & Peters, A. (2017). Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. *Science of the Total Environment*, 579, 1531-1540.

Wu, J., Li, J., Peng, J., Li, W., Xu, G., & Dong, C. (2015). Applying land use regression model to estimate spatial variation of $PM_{2.5}$ in Beijing, China. *Environmental Science and Pollution Research*, 22(9), 7045-7061.

Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, *75*, 16-25.

Xu, W. A., & Yang, L. (2019). Evaluating the urban land use plan with transit accessibility. *Sustainable Cities and Society*, 45, 474-485.

Xu, P., Wang, W., Ji, J., & Yao, S. (2014). Analysis of the contribution of the road traffic industry to the $PM_{2.5}$ emission for different land-use types. *Computational Intelligence and Neuroscience*, 2014, 35.

Yan, D., Lei, Y., Shi, Y., Zhu, Q., Li, L., & Zhang, Z. (2018). Evolution of the spatiotemporal pattern of $PM_{2.5}$ concentrations in China–a case study from the Beijing-Tianjin-Hebei region. *Atmospheric Environment*, 183, 225-233.

Yang, D., Ye, C., Wang, X., Lu, D., Xu, J., & Yang, H. (2018). Global distribution and evolvement of urbanization and $PM_{2.5}$ (1998–2015). *Atmospheric Environment*, 182, 171-178.

Yang, H., Zhang, Y., Zhong, L., Zhang, X., & Ling, Z. (2020). Exploring spatial variation of bike sharing trip production and attraction: A study based on Chicago's Divvy system. *Applied Geography*, *115*, 102130.

Yang, L., Chau, K. W., & Chu, X. (2019). Accessibility-based premiums and proximity-induced discounts stemming from bus rapid transit in China: empirical evidence and policy implications. *Sustainable Cities and Society*, 48, 101561.

Yang, X., Wang, S., Zhang, W., Zhan, D., & Li, J. (2017a). The impact of anthropogenic emissions and meteorological conditions on the spatial variation of ambient SO_2 concentrations: a panel study of 113 Chinese cities. *Science of the Total Environment*, 584, 318-328.

Yang, Q., Yuan, Q., Li, T., Shen, H., & Zhang, L. (2017b). The relationships between $PM_{2.5}$ and

meteorological factors in China: seasonal and regional variations. *International Journal of Environmental Research and Public Health*, 14(12), 1510.

Yang, X., Zheng, Y., Geng, G., Liu, H., Man, H., Lv, Z., He, K., & de Hoogh, K. (2017c). Development of $PM_{2.5}$ and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environmental Pollution*, 226, 143-153.

Yao, L., Yang, L., Yuan, Q., Yan, C., Dong, C., Meng, C., Sui, X., Yang, F., Lu, Y., & Wang, W. (2016). Sources apportionment of $PM_{2.5}$ in a background site in the North China Plain. *Science of the Total Environment*, 541, 590-598.

Yin, S., Wang, X., Xiao, Y., Tani, H., Zhong, G., & Sun, Z. (2017). Study on spatial distribution of crop residue burning and $PM_{2.5}$ change in China. *Environmental Pollution*, 220, 204-221.

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, 145-154.

Zhang, Z., Zhang, X., Gong, D., Quan, W., Zhao, X., Ma, Z., & Kim, S.J. (2015). Evolution of surface O_3 and $PM_{2.5}$ concentrations and their relationships with meteorological conditions over the last decade in Beijing. *Atmospheric Environment*, 108, 67-75.

Zhang, T., Zang, L., Wan, Y., Wang, W., & Zhang, Y. (2019). Ground-level $PM_{2.5}$ estimation over urban agglomerations in China with high spatiotemporal resolution based on Himawari-8. *Science of the Total Environment*, 676, 535-544.

Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period $PM_{2.5}$ prediction based on multivariate linear regression model. *Plos One*, 13(7), e0201011.

Zhao, R., Zhang, Y., & Guo, S. (2019). Construction of an Improved Air Quality Index: a Case Report. *Iranian Journal of Public Health*, *48*(8), 1523-1527.

Zheng, Y., Zhang, Q., Liu, Y., Geng, G., & He, K. (2016). Estimating ground-level $PM_{2.5}$ concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. *Atmospheric Environment*, 124, 232-242.

Zhou, Q., Jiang, H., Wang, J., & Zhou, J. (2014). A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Science of the Total Environment*, 496, 264-274.

Zou, B., Luo, Y., Wan, N., Zheng, Z., Sternberg, T., & Liao, Y. (2015). Performance comparison of LUR and OK in $PM_{2.5}$ concentration mapping: a multidimensional perspective. *Scientific Reports*, 5, 8698.

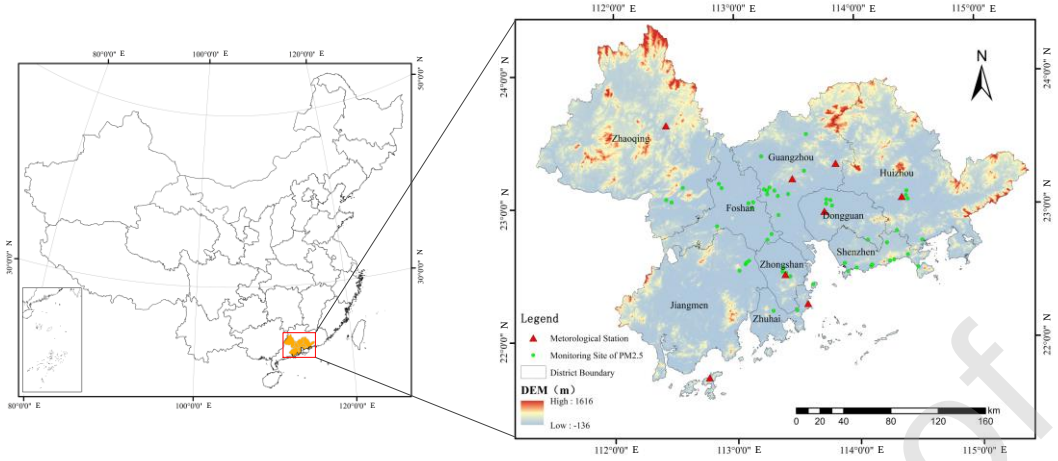


Fig. 1. Location of the study area and spatial distribution of $PM_{2.5}$ monitoring sites and meteorological stations.

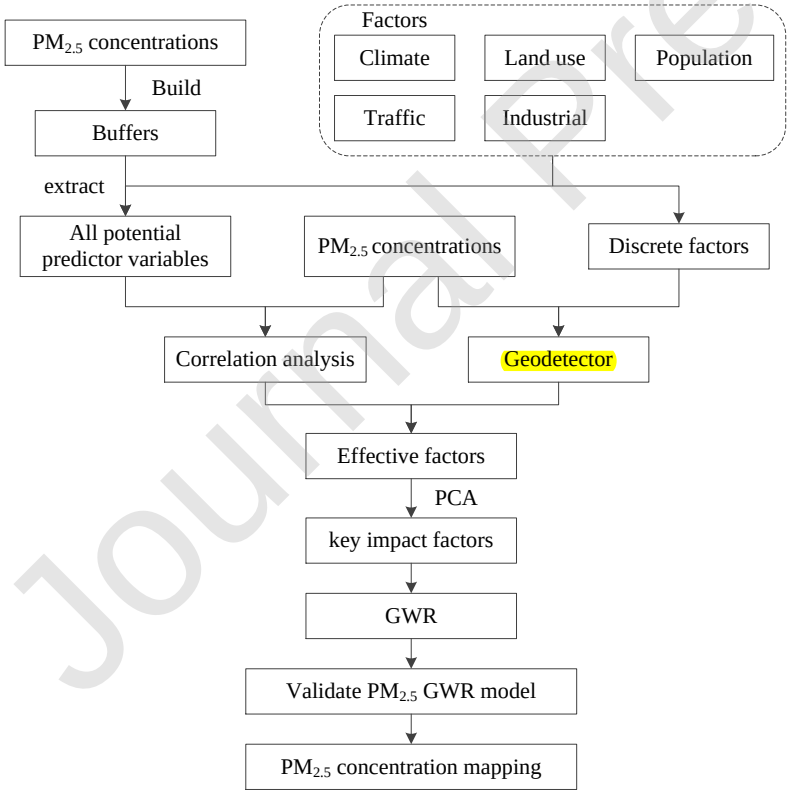


Fig. 2. GWR modeling approach based on Geodetector and PCA.

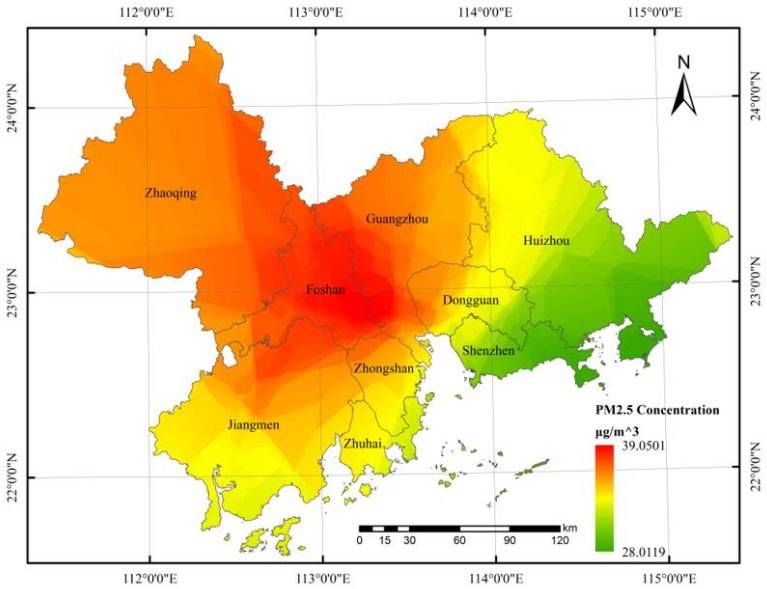


Fig. 3. Spatial distribution of $PM_{2.5}$ concentrations based on an augmented GWR model.

Table 1 Description of potential $PM_{2.5}$ predictor variables used in this study.

Category	Factor/Variable	Unit	Buffer size (km)
Meteorological	Temperature	℃	NA
	Precipitation	mm	NA
	Average wind velocity	m/s	NA
	Maximum wind speed	m/s	NA
	Relative humidity	RH%	NA
	Atmospheric pressure	hPa	NA
	Vapor pressure	hPa	NA
Land use	Agricultural land	m^2	1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10
	Green space	m^2	1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10
	Building land	m^2	1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10
	Water body	m^2	1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10
Traffic	Length of roads	m	0.1, 0.2, 0.3, 0.4, .5, 0.6, 0.7, 0.8, 1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10
Industrial	Number of enterprises	count	1, 3, 5, 7, 10
Demographic	Population	person/ km^2	1, 1.2, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10

Table 2 Relative errors of interpolation using the meteorological data.

Meteorological data / Interpolation method	Temperature	Precipitation	Average wind velocity	Atmospheric pressure	Vapor pressure	Relative humidity	Maximum wind speed
IDW interpolation	2.07%	10.70%	10.52 %	0.65%	2.23%	3.40%	10.85%
Kriging interpolation	2.15%	11.70%	11.60 %	0.88%	2.50%	3.38%	11.04%
Natural neighbor interpolation	2.19%	9.36%	10.94 %	0.74%	2.67%	3.42%	10.92%
Trend interpolation	1.88%	11.99%	14.57 %	0.61%	2.56%	3.45%	12.22%
Spline interpolation	3.06%	11.89%	13.38 %	0.96%	3.40%	3.56%	14.48%

Table 3 Correlations between variables from Geodetector analysis

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Atmospheric pressure (V1)	0.378	–	–	–	–	–	–	–	–	–	–
Temperature (V2)	0.682	0.365	–	–	–	–	–	–	–	–	–
Precipitation (V3)	0.678	0.494	0.273	–	–	–	–	–	–	–	–
Relative humidity (V4)	0.696	0.530	0.528	0.256	–	–	–	–	–	–	–
Vapor pressure (V5)	0.607	0.527	0.537	0.495	0.159	–	–	–	–	–	–
Average wind velocity (V6)	0.725	0.611	0.622	0.664	0.600	0.393	–	–	–	–	–
Maximum wind speed (V7)	0.539	0.697	0.539	0.642	0.625	0.718	0.320	–	–	–	–
Population (V8)	0.696	0.646	0.624	0.522	0.677	0.664	0.628	0.206	–	–	–
Length of roads (V9)	0.644	0.708	0.470	0.525	0.506	0.617	0.591	0.496	0.085	–	–
Land use (V10)	0.621	0.503	0.487	0.445	0.452	0.540	0.513	0.408	0.304	0.085	–
Industrial emissions (V11)	0.468	0.424	0.358	0.329	0.253	0.445	0.442	0.304	0.266	0.204	0.004

Table 4 Details of the effective variables

Variable	Correlation	VIF
Temperature	-0.347**	22.562
Precipitation	0.301**	13.903
Mean wind speed	-0.158	36.899
Maximum wind speed	-0.028	12.159
Relative humidity	-0.169	7.234
Atmospheric pressure	-0.280**	10.402
Vapor pressure	-0.343**	16.346
Agricultural land area in the buffer of 8 km	0.162	3.872
Green space area in the buffer of 4 km	-0.287**	3.288
Building land area in the buffer of 8 km	0.420**	4.366
Water body area in the buffer of 10 km	0.236*	1.484
Length of roads in the buffer of 10 km	0.177	3.284
Number of enterprises in the buffer of 3 km	0.076	1.331
Population counts in the buffer of 5 km	0.184	2.478

*Significant at the 5% level (two-tailed).

**Significant at the 1% level (two-tailed).

Table 5 Details of the PCs.

PCs	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
VIF	1.008	1.007	1.002	1.001	1.001	1.004	1.001	1.001	1.001	1.001	1.001	1.003	1.007	1.011
Contribution	0.342	0.221	0.114	0.093	0.077	0.053	0.031	0.021	0.019	0.014	0.007	0.003	0.002	0.001

Table 6 Accuracy assessment for the augmented GWR model.

Model	R^2	MAE	MRE	RMSE
Augmented GWR	0.84	2.49	0.07	2.94

Table 7 Comparison of different regression models.

Model	Predictor variable	AICc	R^2	Adjusted R^2
OLS	PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8	290	0.7	0.65
GWR	Numerous variables	261	0.81	0.78
The augmented GWR	PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8	273	0.84	0.77