International Journal of Geo-Information

MDPI

Article

A Generalized Additive Model Combining Principal Component Analysis for $PM_{2.5}$ Concentration Estimation

Shuang Li [1,2], Liang Zhai [2,*], Bin Zou [3], Huiyong Sang [2] and Xin Fang [3]

1 College of Geomatics, Shandong University of Science and Technology, Qingdao 266590, China; ls02020029@163.com

2 National Geographic Conditions Monitoring Research Center, Chinese Academy of Surveying and Mapping, Beijing 100830, China; huiyong.sang@casm.ac.cn

3 School of Geosciences and Info-Physics, Central South University, Hunan 410083, China; 210010@csu.edu.cn (B.Z.); xinfang@csu.edu.cn (X.F.)

* Correspondence: zhailiang@casm.ac.cn; Tel.: +86-135-5235-8197

Received: 25 June 2017; Accepted: 10 August 2017; Published: 13 August 2017

Abstract: As an extension of the traditional Land Use Regression (LUR) modelling, the generalized additive model (GAM) was developed in recent years to explore the non-linear relationships between $PM_{2.5}$ concentrations and the factors impacting it. However, these studies did not consider the loss of information regarding predictor variables. To address this challenge, a generalized additive model combining principal component analysis (PCA–GAM) was proposed to estimate $PM_{2.5}$ concentrations in this study. The reliability of PCA–GAM for estimating $PM_{2.5}$ concentrations was tested in the Beijing-Tianjin-Hebei (BTH) region over a one-year period as a case study. The results showed that PCA–GAM outperforms traditional LUR modelling with relatively higher adjusted R^2 (0.94) and lower RMSE (4.08 $\mu g/m^3$). The CV-adjusted R^2 (0.92) is high and close to the model-adjusted R^2, proving the robustness of the PCA–GAM model. The PCA–GAM model enhances $PM_{2.5}$ estimate accuracy by improving the usage of the effective predictor variables. Therefore, it can be concluded that PCA–GAM is a promising method for air pollution mapping and could be useful for decision makers taking a series of measures to combat air pollution.

Keywords: PCA; GAM; $PM_{2.5}$ concentrations; effective predictor variables; utilization rate

1. Introduction

Fine particulate matter consists of particles less than 2.5 μm ($PM_{2.5}$) that are suspended in the atmosphere in solid or liquid form [1]. Due to its potential threat to human health and the environment, $PM_{2.5}$ has been given high priority in research activities in the fields of air pollution and environmental health [2–4]. Recent epidemiological studies have shown an association between $PM_{2.5}$ and adverse effects on human health, including an increased risk of cardiovascular diseases [5,6], heart problems and lung cancer [7], and a significantly reduced birth rate [8,9]. $PM_{2.5}$ pollution has become one of the critical air problems and seriously affects people's daily lives worldwide. Together with rapid economic development and urbanization, heavy air pollution poses serious challenges to environmental sustainability in China. $PM_{2.5}$ pollution in China has become a social problem and has attracted significant attention from the public and government officers. Therefore, a clear and correct understanding of the spatial-temporal characteristics of $PM_{2.5}$ distribution can help us obtain the $PM_{2.5}$ pollution level in different regions, and provide a scientific support for joint prevention and control of $PM_{2.5}$ pollution.

The tapered element oscillating microbalance method (TEOM), is currently considered the most reliable way to collect $PM_{2.5}$ concentrations through ground-level measured $PM_{2.5}$ concentrations [10].

However, the sparseness of monitoring stations cannot meet the urgent need of obtaining the $PM_{2.5}$ concentrations over a large area. Among the available estimation methods, Land Use Regression (LUR) modelling [11] is one of the best approaches whose strengths include large-scale air quality and continuous space estimation. It is an efficient statistical regression model that estimates air pollution by using ground-level monitoring data as the dependent variable, and uses surrounding land use, Aerosol Optical Depth (AOD), meteorological and other auxiliary data as the independent variables [12]. In recent decades, LUR modelling has been widely used to study the spatial distribution of air pollutants, such as $PM_{2.5}$ [13–15], PM_{10} [16,17], NO_2 [17,18], NO_X [18], SO_2 [19], and O_3 [20]. However, most of them depend on presumed linear relationships between the ground-level measured $PM_{2.5}$ concentrations and the independent variables, despite the fact that the linear influencing mechanism on $PM_{2.5}$ concentration is not always suitable for all independent variables. Focusing on this issue, the generalized additive model (GAM) was introduced to capture the non-linear and non-monotonic relationships between variables in a few studies [21–24]. Results of those studies proved that the GAM is effective at identifying the effect of different factors on regional $PM_{2.5}$ concentrations, meaning GAM modelling is a robust method for estimating $PM_{2.5}$ concentrations. Furthermore, having the capacity to integrate linear and non-linear statistical modelling techniques, GAM modelling outperforms traditional ordinary least square (OLS) modelling.

Although those studies addressed the specific issue, some challenges still remain for the regression modelling community regarding the application of LUR models. One of the most important challenges is that these studies ignored the loss of predictor variable information. In the process of modelling, a majority of the effective predictor variables were removed despite their significant correlation with $PM_{2.5}$ concentration. That is to say, all effective predictor variables, which are significantly related to $PM_{2.5}$ concentration, cannot be used in the final regression model. Too many predictor variables can cause an over-fitting problem in regression models, while using an appropriate number of predictor variables may lead to a loss of information significantly related to $PM_{2.5}$ concentration. According to He's research (2017) [24], the variance explained decreased from 75.5% to 73.9% after removing some of the effective predictor variables including PRS (pressure), TEM (temperature), and SSD (sunshine duration). The results suggested that the contributing strength of related influencing factors to the final regression model decreased due to removing some effective predictor variables.

In response to the above challenges, principal component analysis (PCA) is employed to improve the utilization rate of effective predictor variables in this paper. As a basic mathematical analysis method, PCA is the simplest method for eigenfactor-based multivariate analyses. It is used to reduce the number of predictor variables and transform information into new variables that are mutually orthogonal, or uncorrelated, as well as to determine the dominant multivariate relationships [25,26]. PCA is able to remove redundant information among variables by eliminating the collinearity problem and integrating the same variable information together [27,28]. As a commonly used multivariate analysis method, PCA has been gradually applied in air quality studies to analyze voluminous environmental data.

In this study, a generalized additive model combined with principal component analysis (PCA–GAM) is proposed to estimate $PM_{2.5}$ concentrations over a large area. PCA was used to simplify the complexity of relationships among variables to improve the utilization rate of effective predictor variables. Moreover, GAM was used to explore the linear or non-linear relationships between $PM_{2.5}$ concentrations and the independent variables. In order to quantitatively evaluate the performance of the PCA–GAM model, the Beijing-Tianjin-Hebei (BTH) region was used as the case study for estimating $PM_{2.5}$ concentrations. The research results will help promote the reliability and stability of the LUR method for large area $PM_{2.5}$ mapping and provide support for decision makers when seeking comprehensive environmental improvement.

2. Experiments

2.1. Study Area and Data Collection

The BTH region is located in Northern China, and includes Beijing, Tianjin and the province of Hebei, covering an area of 216,957 km^2. The latitude and longitude of the BTH region are 113°04′ to 119°53′ and 36°01′ to 42°37′. Because of the rapid industrial and economic development, a long history of human occupation, and the terrain conditions, the BTH region has become one of the typical urban pollution areas, with a stable geographic and meteorologic environment, high air pollution levels, and has some of the most intensive urban $PM_{2.5}$ monitoring sites in the world. To persistently and effectively control $PM_{2.5}$ pollution, the government has taken a series of measures, for example, enacting vehicle use restrictions and has closed several polluting industrial plants.

According to the previous LUR research findings on the selection of geographical feature characteristics [14,29–31], data collected for LUR modelling in this study contains annual average $PM_{2.5}$ concentrations, elevation, AOD, climate characteristics (temperature, wind speed, relative humidity, atmospheric pressure, and precipitation), road traffic, land use and cover, industrial plants, and surface dust. The distribution of $PM_{2.5}$ monitoring sites and the partial basic geographical feature data, within the BTH region during the study period of 1 January 2015 to 31 December 2015, are shown in Figure 1. Seventy-eight $PM_{2.5}$ monitoring sites located in the BTH region are all urban sites. Owing to the high cost of in situ observation, ground stationary monitoring networks are ordinarily sparse or even unavailable. As a result, the ground-level $PM_{2.5}$ concentrations were used as the dependent variable, and surrounding land use, transportation, meteorological, and other auxiliary data are fully used as the predictive variables to establish a regression model to make up for the sparse air quality monitoring sites in this study.

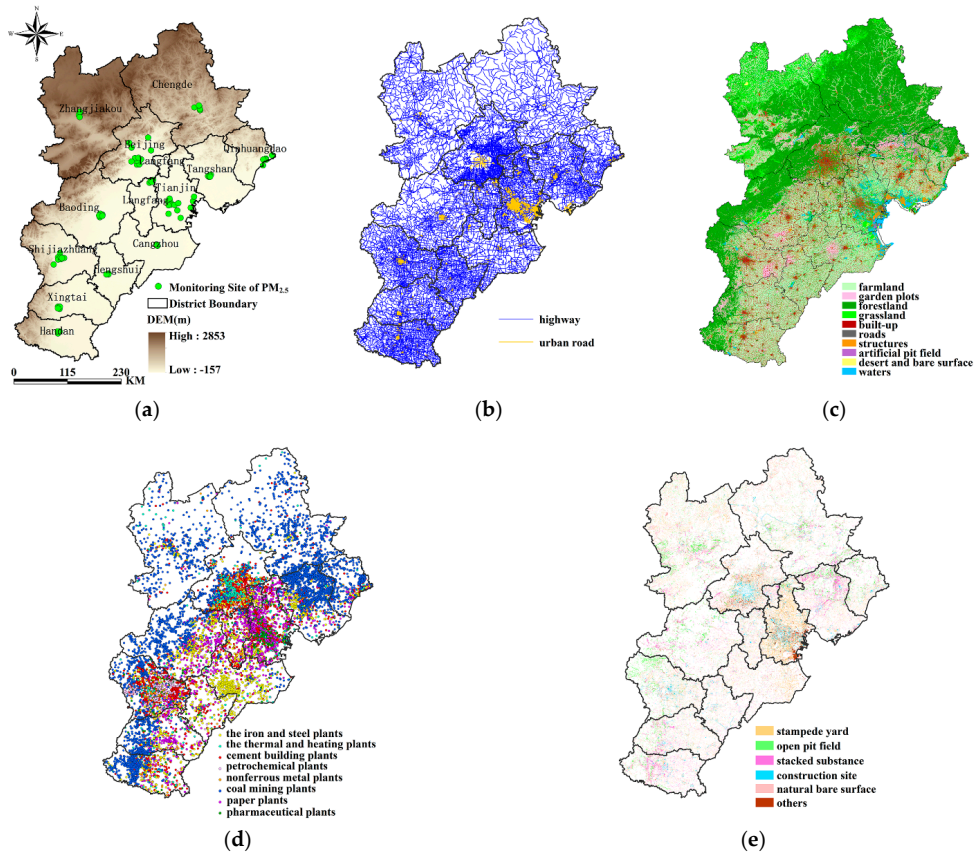

Figure 1. Study area and partial basic geographical feature data: (**a**) $PM_{2.5}$ monitoring sites and elevation; (**b**) road traffic; (**c**) land use/cover; (**d**) industrial plants; and (**e**) surface dust.

2.2. Methods

The methodology in this study includes four parts: Predictor variables extraction and screening, regression modelling, model validation, and $PM_{2.5}$ concentrations mapping. The framework of the study procedure is shown in Figure 2.

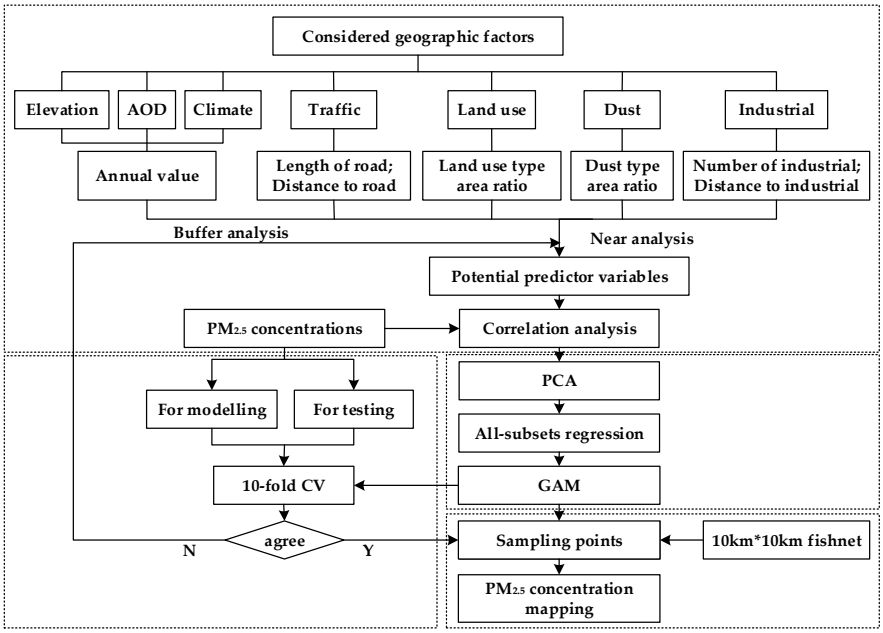

Figure 2. Framework of the study procedure.

2.2.1. Predictor Variable Extraction and Screening

Based on the previous LUR research [10,17,18,32], the potential predictor variables used in this study included the area ratio of each type of land use, the area ratio of surface dust, the length of road, and the number of industrial plants. The characteristic values were extracted at a 100–10,000 m (100, 200, 400, 500, 600, 800, 1000, 2000, 3000, 4000, 5000, 6000, 8000, and 10,000 m) buffering radius based on previous findings and experiments [10,14,29]. Moreover, the distance to a nearest road or industrial plant, elevation, the annual averages of AOD, as well as the annual averages of climate characteristics, were also included as the potential predictor variables in this study. After all the potential predictor variables were extracted in ArcGIS 10.2, the Pearson's correlation analysis [33] and two tailed significance test ($\alpha = 0.05$) were conducted to screen out effective predictor variables that are most closely related to $PM_{2.5}$ concentration, using IBM SPSS Statistics 22.0 software. Furthermore, for the remaining predictor variables, many buffers existed for one type of predictor variable; however, the performance of the predictor variable was affected by the buffer. Thus, the optimal spatial scale of one type of effective predictor variable was chosen by the criterion of the highest Pearson's correlation coefficient amongst all of the buffers.

2.2.2. Regression Modelling

The regression modelling, which is composed of PCA, all-subsets regression, and GAM, played an important role in this study.

1. PCA refers to a mathematical method that transforms the original set of inter-correlated variables into a new set of an equal number of independent uncorrelated variables, which gives the linear combination of the original set of data. It maximizes the correlation between the original variables to form new variables that are mutually orthogonal, or uncorrelated [26]. The new variables are ordered in such a way that the first new variable explains most of the variance in the data, and

each subsequent one accounts for the largest proportion of variability that has not been accounted for by its predecessors. For this process, PCA was employed to transform the final inter-correlated effective predictor variables into principal components (PCs) in PAST software. Additionally, the redundancy information of effective predictor variables was removed by using the PCs instead of the original explanatory variables, and integrating the same variable information together.

2. To avoid the over-fitting problem caused by too many PCs in the regression modelling, we used all-subsets regression to select the optimal subset of variables in R Studio. As one of the most common methods for selecting the final predictor variables from too many variables, all-subsets regression tests all possible subsets of the set of potential independent variables [34]. If there are k potential independent variables besides the constant, then there are 2^k distinct subsets of them to be tested, including the empty set which corresponds to the mean model [35,36]. Several measures with respect to the selection criteria have been proposed, such as the adjusted coefficient of determination (adjusted R^2), Mallow's Cp, and the Akaike Information Criterion (AIC) [37]. The adjusted R^2 was used as the selection criteria to select the optimal subset of PCs in this study.
3. After the pre-screening of multivariate variables, the package of "mgcv" in R Studio was used to fit GAM as implemented by the gam() function, which generalized multivariate regression by relaxing the assumptions of linearity and normality, replacing regression lines by smooth lines [38]. In this process, the linear or non-linear relationships between $PM_{2.5}$ concentration with associated contributing factors, were fitted with thin plate regression splines by using the "GCV" method to automatically choose a smoothing parameter [39]. The one degree of freedom indicated that the predictor variable was fitted with a parametric linear term rather than a smoothed term. The finalized regression model presented in this article was determined such that the model AIC value is among the lowest of all the models [40]. Additionally, a significant test was also employed using the 0.05 level to check whether each term remaining in the finalized model was statistically significant [22].

2.2.3. Model Validation

The model performance of PCA–GAM was validated by comparing it with OLS and GAM based on the corresponding domain date sets. The adjusted R^2, AIC, root mean square error (RMSE), mean percentage error (MPE), as well as the mean absolute percentage error (MAPE) were used as the statistics to evaluate the prediction ability and reliability of the three models. As a general rule, a higher adjusted R^2 and smaller AIC, RMSE, and MPE mean the model is more perfectible. At the same time, the 10-fold cross validation (10-fold CV) [41,42] was employed to test the feasibility and robustness of the model. In this process, the dataset was randomly divided into 10 folds, among which 9 folds were selected as the training set and the remaining fold was used as the test set. This progress was repeated 10 times until all samples were tested.

2.2.4. $PM_{2.5}$ Concentration Mapping

To visualize the spatial distribution of annual $PM_{2.5}$ concentrations in the study area, we created a fishnet with a resolution of 10 km × 10 km to obtain the sampling points in the BTH region. $PM_{2.5}$ concentrations were then predicted using PCA–GAM. Finally, the continuous raster surfaces of annual $PM_{2.5}$ concentrations were produced though the Ordinary Kriging (OK) method, which weighted the surrounding measured values to derive a prediction for an unmeasured location, not only based on the distance but also on the overall spatial autocorrelation of the measured point [12].

3. Results

3.1. Descriptive Effective Predictor Variables

The Pearson's correlation was performed between all the potential predictor variables and annual $PM_{2.5}$ concentrations. With a two-tailed significance of less than 0.05, 17 effective predictor variables

were screened: AOD, Temp (temperature), PS (pressure), PE (precipitation), RH (relative humidity), Elev (elevation), N $Road_{all}$ (distance to the nearest of all roads), $N\_Industry_{205}$ (distance to the nearest petrochemical plant), $Industry_{208\_3000m}$ (number of paper plants in a buffer of 3000 m), $Dust_{0718\_1000m}$ (the area ratio of stampede yard in a buffer of 1000 m), $Dust_{0810\_5000m}$ (the area ratio of open pit field in a buffer of 5000 m), $Cover_{1\_8000m}$ (the area ratio of farmland in a buffer of 8000 m) , $Cover_{3\_8000m}$ (the area ratio of grassland in a buffer of 8000 m), $Cover_{5\_8000m}$ (the area ratio of built-up land in a buffer of 8000 m), $Cover_{6\_8000m}$ (the area ratio of roads in a buffer of 8000 m), $Cover_{8\_8000m}$ (the area ratio of artificial pit fields in a buffer of 8000 m), $Cover_{9\_8000m}$ (the area ratio of desert and bare surface in a buffer of 8000 m). The histograms of effective predictor variables are illustrated in Figure 3, which shows that all the variables are roughly unimodal and log-normally distributed. It is easy to find that all the effective predictor variables have similar distributions of $PM_{2.5}$ concentrations. The overall mean and the standard deviation value of the $PM_{2.5}$ concentrations at the monitoring sites in the BTH region are 76.505 $\mu g/m^3$ and 20.445 $\mu g/m^3$, respectively. The maximum, minimum, mean, and standard deviation for all the effective predictor variables are also presented in Figure 3. All of these values show the range and fluctuation of the effective predicted variables, which, from another perspective, reflect the complexity of the effective predictor variables associated with the $PM_{2.5}$ concentration.

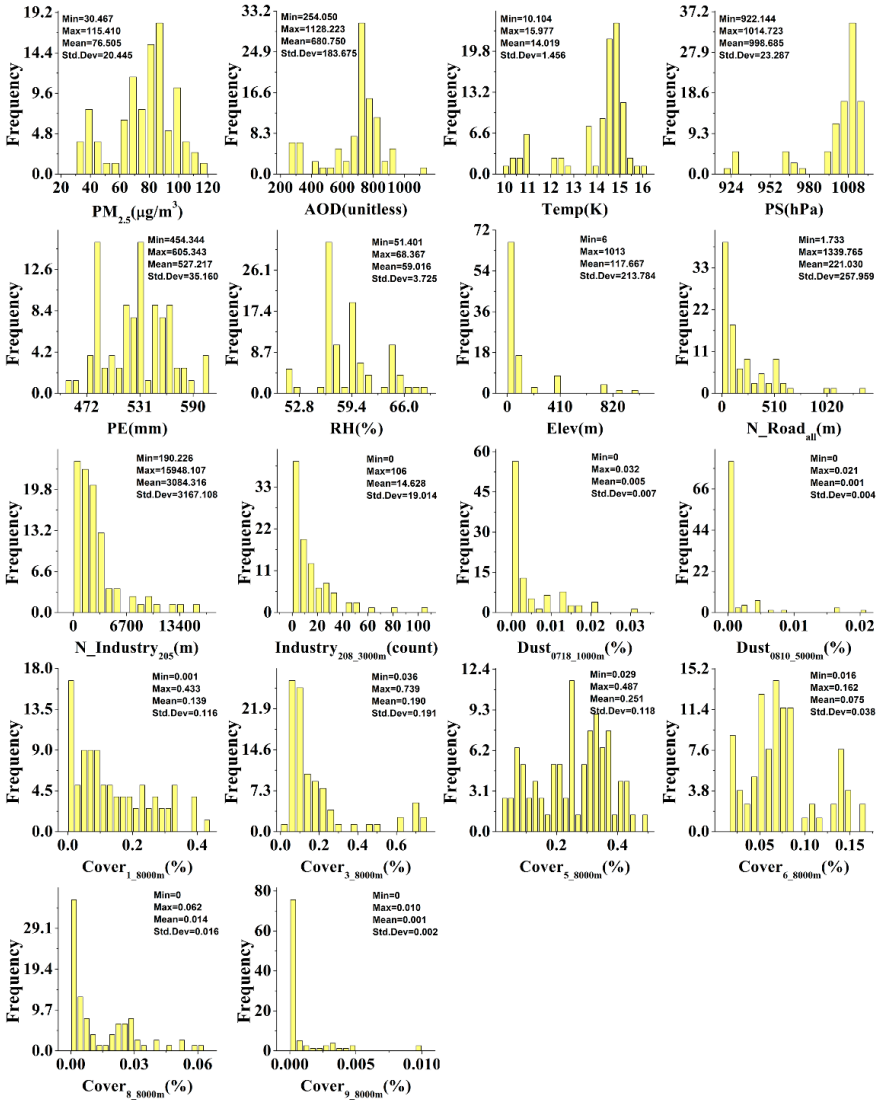

Figure 3. Histogram and description of $PM_{2.5}$ concentrations and potential predictor variables.

3.2. Model Fitting and Validation

The PCA was used to transform the original set of effective predictor variables into a new set of an equal number of PCs. The transformed explanatory variables were labeled PC1 to PC17, which were conducted to select the optimal subset of explanatory variables. Figure 4, which was drawn in R Studio, shows the results of all-subsets regression. With regard to Figure 4, the *y*-axis represents the adjusted R^2, each row represents one model, and colored rectangles represent the explanatory variables which were included in each model. Figure 4 shows that there are four candidate models with the highest adjusted R^2 with a value of 0.89. According to the AIC criterion, the candidate models with the lowest AIC value was used, and had independent variables were PC1, PC2, PC4, PC5, PC6, PC8, and PC17. The AIC value and variance explained are 517.10 and 92.5%, respectively. Additional, from the results of the two-tailed significance test ($\alpha = 0.05$), it was found that PC6 does not significantly influence $PM_{2.5}$ concentration. All the remaining independent variables reached a significant level when PC6 was excluded from the model. Therefore, PC1, PC2, PC4, PC5, PC8, and PC17 were determined as independent variables in the finalized regression model. The value of variance explained for the finalized regression model is 96.00%.

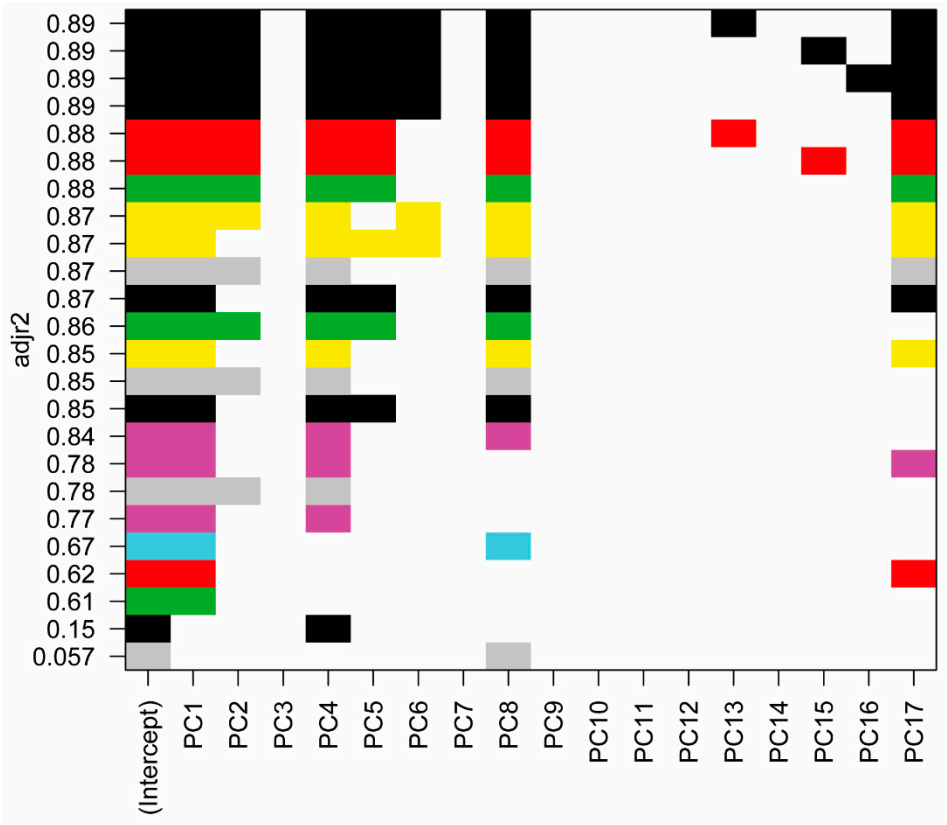

Figure 4. All-subsets regression.

A comparison of performance among three models are presented in Table 1. It can be seen that the independent variables are composed of PE, AOD, $Cover_{1\_8000m}$, $Cover_{3\_8000m}$, $Cover_{8\_8000m}$ for OLS or GAM, and PC1, PC2, PC4, PC5, PC8, PC17 for PCA–GAM, respectively. The adjusted R^2 values of OLS, GAM and PCA–GAM are 0.83, 0.90 and 0.94, respectively, the 10-fold CV adjusted R^2 value of three models are 0.83, 0.92 and 0.92, respectively. Other accuracy indicators including AIC, RMSE, MPE, and MAPE of the PCA–GAM model were 495.52, 4.08 $\mu g/m^3$, −0.39%, and 4.10%, respectively, which are significantly less than OLS (563.37, 8.10 $\mu g/m^3$, −1.39% and 8.63%), or GAM (528.23, 5.50 $\mu g/m^3$, −0.72%, and 5.78%). To further comprehensively compare the performance of the OLS, GAM, and PCA–GAM models, scatter plots between the observed and estimated values of fitting and validating

results for these three types of models are demonstrated in Figure 5. For model fitting, the adjusted R^2 value, computed from PCA–GAM, is 0.96, higher than that computed using OLS (0.84) or GAM (0.93). This is also true in model validating; while the PCA–GAM had the highest adjusted R^2 (0.77) and lowest RMSE (9.88 μg/m^3) among the three models.

Table 1. The regression results of three models.

Model	Independent Variables	Adj_R^2	AIC	RMSE (μg/m^3)	MPE (%)	MAPE (%)	CV Adj_R^2
OLS	PE, AOD, $Cover_{1\_8000m}$, $Cover_{3\_8000m}$, $Cover_{8\_8000m}$	0.83	563.37	8.10	−1.39	8.63	0.83
GAM	PE, AOD, $Cover_{1\_8000m}$, $Cover_{3\_8000m}$, $Cover_{8\_8000m}$	0.90	528.23	5.50	−0.72	5.78	0.92
PCA–GAM	PC1, PC2, PC4, PC5, PC8, PC17	0.94	495.52	4.08	−0.39	4.10	0.92

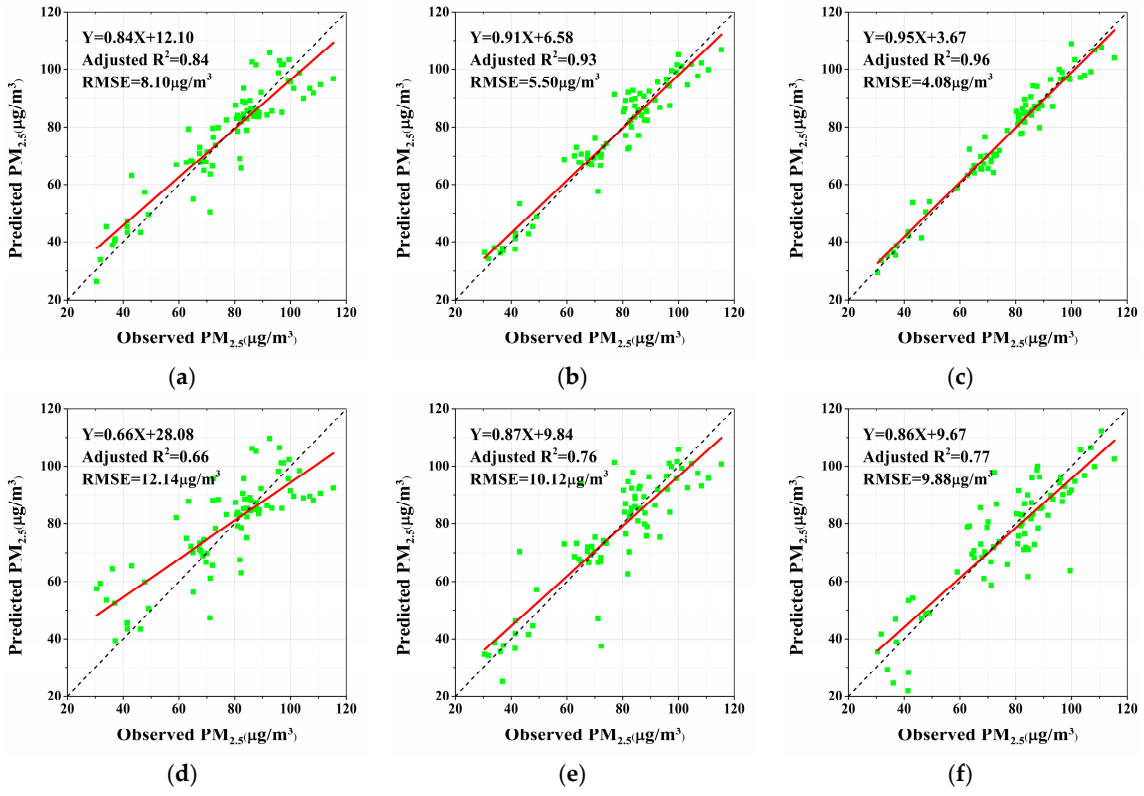

Figure 5. Scatter plots of fitting and validating results for three models. (**a**) OLS model fitting results; (**b**) GAM model fitting results; (**c**) PCA–GAM model fitting results; (**d**) OLS model validating results; (**e**) GAM model validating results; and (**f**) PCA–GAM model validating results.

Figure 6 shows the fitted curve of each independent variable in PCA–GAM model. It can be observed that the greater the degree of freedom, the more fluctuation in the fitting curve. Among all the fitted curves, only the independent variable PC2 has a straight fitted line corresponding to the one degree of freedom. For all others, there is a non-linear relationship between independent variables and $PM_{2.5}$ concentrations. Besides, these relationships varied among independent variables. The different fitted curves are shown in Figure 6. The relationship between PC1 and $PM_{2.5}$ concentration is monotonically increases, while $PM_{2.5}$ concentration decreases monotonically with the increase of PC4 or PC5. Moreover, there are fluctuating changes in the effects of PC8 and PC17 on $PM_{2.5}$ concentration. The results show that introducing PCA into the GAM model do not weaken the advantages of GAM,

which is the ability to capture the highly non-linear and non-monotonic relationship between variables. Considering the utilization rate of effective predictor variables, PCA–GAM still can successfully capture the linear and non-linear relationship between $PM_{2.5}$ variation and the associated contributing factors.

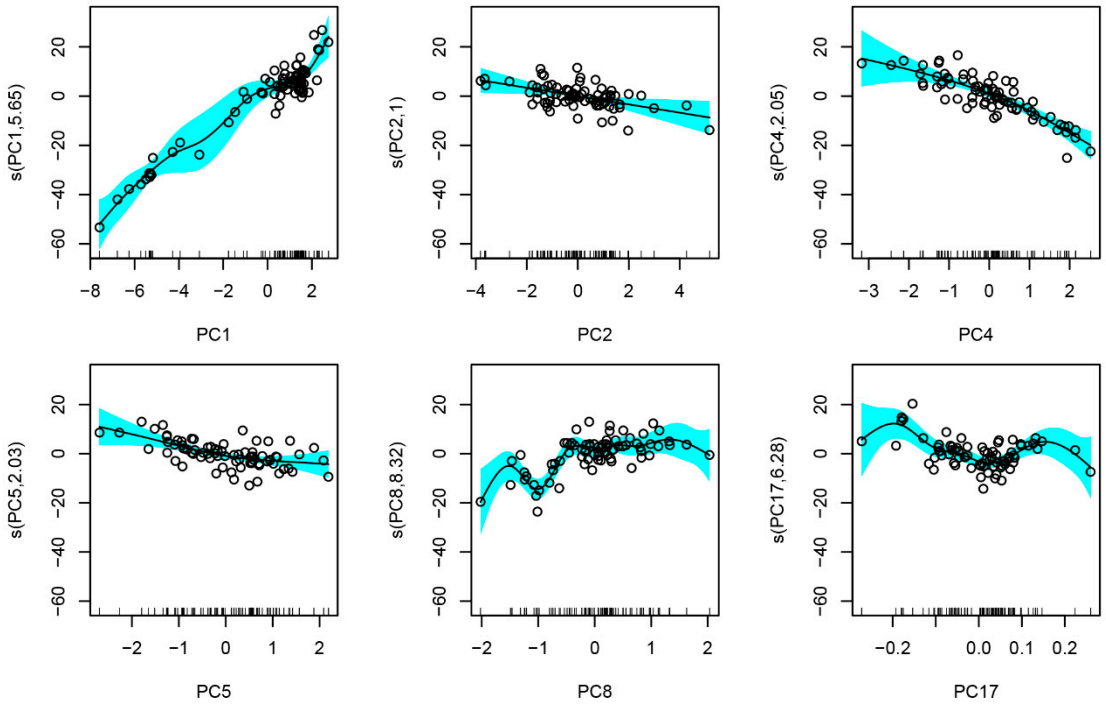

Figure 6. Results of PCA–GAM analysis. The *x*-axis represents the frequency of data; the *y*-axis represents the smooth fitted values. Degrees of freedom for linear or non-linear fits are in the parentheses on the *y*-axis. The solid line represents the fitted curve of each independent variable, the blue area represents the 95% confidence interval.

3.3. $PM_{2.5}$ Concentration Mapping

Figure 7 demonstrates the spatial distribution of annual estimated $PM_{2.5}$ concentrations during the study period in the BTH region for PCA–GAM based map. The annual mean $PM_{2.5}$ concentration in the BTH region ranges from 27.54 $\mu g/m^3$ to 118.98 $\mu g/m^3$, and the overall mean value is 80.34 $\mu g/m^3$. The overall observed trend in the spatial patterns of the annual mean $PM_{2.5}$ concentration is the same as in Bin's study [22], proving the reliability of the mapping results when using PCA–GAM. Furthermore, it is obviously that the annual $PM_{2.5}$ concentrations have significant spatial aggregation in this study. Clearly, higher concentrations of estimated $PM_{2.5}$ normally cluster into several cities, while the concentration of estimated $PM_{2.5}$ decreases gradually with the increasing distance from the center of the city. This is a clearly seen in Beijing, Tianjin, Baoding, Shijiazhuang, and Zhangjiakou. As the capital of China, and one of the most $PM_{2.5}$-polluted region in the BTH region, the overall mean value of $PM_{2.5}$ concentration in Beijing is 86.63 $\mu g/m^3$. $PM_{2.5}$ concentrations in the region southeast of Beijing were generally higher than those in the northwest. This mapping result is the same as the previous research on estimating $PM_{2.5}$ concentration in Beijing [1]. Additionally, it is worth noting that the coastal areas, like the eastern part of Cangzhou and the southern part of Tangshan, are conducive to low concentrations of $PM_{2.5}$ due to the meteorological conditions.

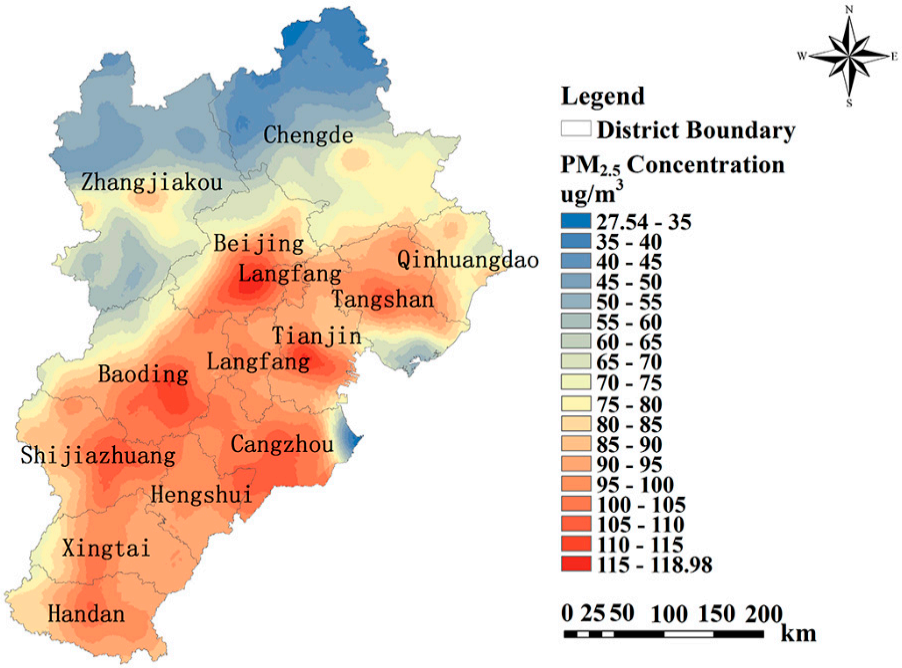

Figure 7. Spatial distributions of estimated $PM_{2.5}$ concentrations.

4. Discussion

With the development of GIS technology, LUR models have been increasingly used for studying the spatial distribution of environmental pollutants. As an extension to the traditional LUR model, GAM was developed to explore the non-linear relationships between air pollutants and the factors impacting air pollution in a few studies [22,23,42]. These studies focused on the improvement of the model and spatio-temporal analysis of air pollutants, but have had a low utilization rate of effective predictor variables. In this study, a PCA–GAM method was proposed to estimate $PM_{2.5}$ concentration in a large area for the first time. The proposed method enriched the limited evidence for developing the LUR model for estimating $PM_{2.5}$ concentration in a large area by taking the utilization rate of effective predictor variables into consideration.

The mechanism between $PM_{2.5}$ concentration and the factors that influence it is rather complex. We considered as many related influencing factors as possible. In this study, there are two advantages in regressing the response variable against PCs rather than directly on the effective predictor variables. Firstly, similar variable information was integrated together and the high correlation among the effective predictor variables was eliminated by removing redundancy information amongst them. Secondly, the final regression model was reasonably and logically consistent with all effective predictor variables that contributed to it.

It is noteworthy that PC6 was removed from the candidate because it did not have a significant influence on $PM_{2.5}$. However, a difference was seen when compared to He's research [24] with the variance explained decreasing from 75.5% to 73.9% after removing some of the effective predictor variables. The variance explained for the regression model in this study actually increased from 92.5% to 96.0%. The main reason was that PC1, PC2, PC4, PC5, PC8, and PC17 were determined as independent variables in PCA–GAM, which implies that all effective predictor variables contribute to the finalized regression model. PC6 might be redundant information that did not contribute to the model. For that, the model could better explain more variability in $PM_{2.5}$ concentration with the removal of the redundant information. In contrast, due to the number of related influencing factors being reduced from directly removing original effective predictor variables, the accuracy of the regression model results will decrease. This conclusion is also supported by the GAM model result in

this paper, with the explained variance decreasing from 96.4% to 92.8% after removing the temperature variable from the model.

According to the case study results in this paper, it was found that PCA–GAM was the best method for estimating $PM_{2.5}$ concentration with the highest adjusted R^2 (0.94) and lowest AIC (495.52) when compared with OLS and GAM using the same datasets. The results also showed that PCA–GAM had the significantly lowest RMSE, MPE, and MAPE, which meant that PCA–GAM can relatively accurately explain more information in $PM_{2.5}$ estimation than the other two models. PCA–GAM has also been proved to have good model reliability and robustness with the large CV adjusted R^2 (0.92). With pairwise comparison of the three models, the prediction accuracy of the PCA–GAM is obviously higher than that of the traditional OLS model. Clearly, the GAM model in this paper also outperformed the OLS model as shown in previous studies by simultaneously considering the linear and non-linear relationships between $PM_{2.5}$ variation and the associated contributing factors. The advantages of PCA can be explored by comparing PCA–GAM with GAM. Although the values of the statistics changed a little, the results showed that the accuracy of the PCA–GAM was indeed improved when compared to GAM alone. Meanwhile, the scatter plots of fitting and validating the results for the three models suggested that PCA–GAM is superior to OLS or GAM. All the results proved the necessity of PCA–GAM modelling in using PCA to simplify the complexity of the relationships among the variables to improve the utilization rate of effective predictor variables. Due to the complex interactions between $PM_{2.5}$ concentration and the contributing factors, a combination of PCA and GAM can effectively improve the performance of $PM_{2.5}$ concentration estimates. In addition, the results in Figure 6 also revealed that the relationships between $PM_{2.5}$ concentration and the independent variables may be linear or non-linear, which highlights that it is essential to integrate linear and non-linear statistical techniques into the LUR model.

Additionally, with the same conclusions as previous studies in the BTH region [22,29,43], the results in this study also demonstrate that the spatial patterns of the annual mean $PM_{2.5}$ concentration have a significant northwest-to-southwest increasing gradient. Dispersion conditions of topography and meteorology could account for these spatial patterns. In this study, the distribution of the annual mean $PM_{2.5}$ concentration had significantly spatial heterogeneity and spatial aggregation, which previous research did not reveal. Higher concentrations of estimated $PM_{2.5}$ normally clustered at several points, corresponding to the rapid economic development, industrial activities, heavy traffic, and high population density in the cities. Additionally, the coastal areas with low concentrations of $PM_{2.5,}$ due to the meteorological conditions, were also reflected in this study using PCA–GAM. In addition to using conventional feature variables, the refined industrial polluting sources and ground dust surfaces were also employed as predictor variables for $PM_{2.5}$ concentration estimation in this study. Considering additional contribution factors, the accuracy of the model proposed was further improved. In summary, with similar or higher estimation accuracy, the PCA–GAM method had better performance in visualized $PM_{2.5}$ concentration mapping than traditional LUR modelling. It can be concluded that the PCA–GAM method could be useful to help scholars estimate the concentrations of air pollutions in other large areas.

According to the comparison of the results of this study, the PCA–GAM method performs better than the traditional LUR model, but the results from this research still has some limitations. Firstly, prevention of air pollution has been an important countermeasure of human sustainable development. In order to help decision makers by providing an overall understanding of the regional variations in $PM_{2.5}$ concentrations, we developed PCA–GAM for estimating $PM_{2.5}$ concentration in a large area over a one–year period. Considering the advantages of PCA–GAM in accurately estimating $PM_{2.5}$ concentration, PCA–GAM could be developed for a shorter time scale, such as seasonal or daily scales. The temporal variations of $PM_{2.5}$ concentration in different time scales should also be investigated in future research. Secondly, some studies suggested that wind direction plays a significant role in the LUR model, because it could affect the dispersion of air pollutants [44]. Thus, the direction of the prevailing winds would be needed as a predictor variable in the regression model. In addition, PCA–GAM,

using globally fixed parameters, assumed that the relationship between $PM_{2.5}$ concentration and independent variables did not vary spatially, which ignores the spatial non-stationarity relationship of environmental variables with air pollution. Therefore, in future work, we could introduce the spatial non-stationarity relationship into PCA–GAM to improve the model performance.

5. Conclusions

In this study, a method of PCA–GAM was proposed for the first time to estimate $PM_{2.5}$ concentration in a large area, to improve the utilization rate of effective predictor variables. The proposed model was validated with a case study of the BTH region over a one-year period. Results of this study indicated that the PCA–GAM model could not only improve the utilization rate of effective predictor variables, but also simultaneously take into account linear and non-linear relationships between $PM_{2.5}$ concentration and the independent variables. The adjusted R^2 (0.94), RMSE (4.08 $\mu g/m^3$), and other accuracy indicators of the case study also indicated the model clearly outperformed those in previously reported studies. Meanwhile, the results of $PM_{2.5}$ concentration mapping accurately reflected the actual sources of serious pollution in the BTH region. As a novel and reliable method, the PCA–GAM model presented in this study provides a general framework for effectively estimating concentrations of air pollution in a large area. It could be a promising way to provide support for air pollution concentrations mapping and helpful to policy makers, environmentalists, and epidemiologists in understanding the complex spatial variations of regional ambient air quality.

Acknowledgments: The research work was supported by the National Geographical Conditions Monitoring Project (B1701), Basic Research Funding in CASM (grant number 7771716), the Program for the 2016 Young Academic and Technological Leaders of NASG, funded by Key Laboratory of Geo-informatics of NASG (Q1702), the Open Fund from the Key Laboratory for National Geographic Census and Monitoring, National Administration of Surveying, Mapping and Geoinformation (2016NGCM ZD03), the Fundamental Research Funds for the Central Universities of Central South University (2016zzts089). We would also like to acknowledge every member of the BTH GCM group from Beijing Institute of Surveying and Mapping, Tianjin Institute of Surveying and Mapping, and Hebei Bureau of Geoinformation. We thank the anonymous reviewers for their helpful comments.

Author Contributions: Liang Zhai and Bin Zou conceived and designed the experiments; Shuang Li performed the experiments; Liang Zhai, Shuang Li, Bin Zou and Huiyong Sang analyzed the data; Shuang Li and Xin Fang contributed reagents/materials/analysis tools; Shuang Li and Liang Zhai wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, L.; Liu, J.; He, Z. Self-Adaptive Revised Land Use Regression Models for Estimating $PM_{2.5}$ Concentrations in Beijing, China. *Sustainability* **2016**, *8*, 786. [CrossRef]
2. Krstic, G. A reanalysis of fine particulate matter air pollution versus life expectancy in the United States. *J. Air Waste Manag. Assoc.* **2012**, *62*, 989–991. [CrossRef] [PubMed]
3. Silva, R.A.; West, J.J.; Zhang, Y.; Anenberg, S.C.; Lamarque, J.F.; Shindell, D.T.; Collins, W.J.; Dalsoren, S.; Faluvegi, G.; Folberth, G.; et al. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environ. Res. Lett.* **2013**, *8*, 034005. [CrossRef]
4. Lim, J.M.; Jeong, J.H.; Lee, J.H.; Moon, J.H.; Chung, Y.S.; Kim, K.H. The analysis of $PM_{2.5}$ and associated elements and their indoor/outdoor pollution status in an urban area. *Indoor Air* **2011**, *21*, 145–155. [CrossRef] [PubMed]
5. Hoek, G.; Krishnan, R.M.; Beelen, R.; Peters, A.; Ostro, B.; Brunekreef, B.; Kaufman, J.D. Long-term air pollution exposure and cardio- respiratory mortality: A review. *Environ. Health* **2013**, *12*, 43. [CrossRef] [PubMed]
6. Giorginia, P.; Di Giosia, P.; Grassi, D.; Rubenfire, M.; Brook, R.D.; Ferri, C. Air pollution exposure and blood pressure: An updated review of the literature. *Curr. Pharm. Des.* **2016**, *22*, 28–51. [CrossRef]

7. Pope, C.A.; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *J. Am. Med. Assoc.* **2002**, *287*, 1132–1141. [CrossRef]
8. Lakshmanan, A.; Chiu, Y.H.M.; Coull, B.A.; Just, A.C.; Maxwell, S.L.; Schwartz, J.; Gryparis, A.; Kloog, I.; Wright, R.J.; Wright, R.O. Associations between prenatal traffic-related air pollution exposure and birth weight: Modification by sex and maternal pre-pregnancy body mass index. *Environ. Res.* **2015**, *137*, 268–277. [CrossRef] [PubMed]
9. Ross, Z.; Ito, K.; Johnson, S.; Yee, M.; Pezeshki, G.; Clougherty, J.E.; Savitz, D.; Matte, T. Spatial and temporal estimation of air pollutants in New York City: Exposure assignment for use in a birth outcomes study. *Environ. Health* **2013**, *12*. [CrossRef] [PubMed]
10. Fang, X.; Zou, B.; Liu, X.; Sternberg, T.; Zhai, L. Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling. *Remote Sens. Environ.* **2016**, *186*, 152–163. [CrossRef]
11. Briggs, D.J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Lebret, E.; Pryl, K.; Van Reeuwijk, H.; Smallbone, K.; Van Der Veen, A. Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 699–718. [CrossRef]
12. Meng, X.; Fu, Q.Y.; Ma, Z.W.; Chen, L.; Zou, B.; Zhang, Y.; Xue, W.B.; Wang, J.N.; Wang, D.F.; Kan, H.D.; et al. Estimating ground-level PM_{10} in a Chinese city by combining satellite data, meteorological information and land use regression model. *Environ. Pollut.* **2015**, *208*, 177–184. [CrossRef] [PubMed]
13. Jiao, L.M.; Xu, G.; Zhao, S.L.; Dong, T.; Li, J.Y. LUR-based simulation of the spatial distribution of $PM_{2.5}$ of Wuhan. *Geomat. Inf. Sci. Wuhan Univ.* **2015**, *40*, 1088–1094.
14. Zhai, L.; Zou, B.; Fang, X.; Luo, Y.; Wan, N.; Li, S. Land Use Regression Modeling of $PM_{2.5}$ Concentrations at Optimized Spatial Scales. *Atmosphere* **2017**, *8*, 1. [CrossRef]
15. Li, J.; Zhai, L.; Sang, H.Y.; Zhang, Y.; Yuan, J. Comparison of different spatial interpolation methods for $PM_{2.5}$. *Sci. Surv. Mapp.* **2016**, *41*, 50–54.
16. Esra, P.; Gunay, S. The Comparision of Partial Least Squares Regression, Principal Component Regression and Ridge Regression with Multiple Line Regression for Predicting PM_{10} Concentration Level Based on Meteorological Parameters. *J. Data Sci.* **2015**, *13*, 663–692.
17. Vienneau, D.; de Hoogh, K.; Bechle, M.J.; Beelen, R.; van Donkelaar, A.; Martin, R.V.; Millet, D.B.; Hoek, G.; Marshall, J.D. Western European land use regression incorporating satellite and ground-based measurements of NO_2 and PM_{10}. *Environ. Sci. Technol.* **2013**, *47*, 68–77. [CrossRef] [PubMed]
18. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.Y.; Künzli, N.; Schikowski, T.; Marcon, A.; et al. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10–23. [CrossRef]
19. Zou, B.; Wilson, J.G.; Zhan, F.B.; Zeng, Y.; Wu, K. Spatial-temporal Variations of Regional Ambient Sulfur Dioxide Concentration and Source Contribution Analysis. *Atmos. Environ.* **2011**, *45*, 4977–4985. [CrossRef]
20. Diem, J.E.; Comrie, A.C. Predictive mapping of air pollution involving sparse spatial observations. *Environ. Pollut.* **2002**, *119*, 99–117. [CrossRef]
21. Hastie, T.; Tibshirani, R. Generalized Additive Models. *Stat. Sci.* **1986**, *1*, 297–318. [CrossRef]
22. Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z. Satellite Based Mapping of Ground $PM_{2.5}$ Concentration Using Generalized Additive Modeling. *Remote Sens.* **2017**, *9*, 1. [CrossRef]
23. Jiao, L.M.; Jin, J.M. Regional $PM_{2.5}$ Concentration Effect Factors Identification and Correlation Analysis Based on GAM. *Environ. Sci. Technol.* **2015**, *38*, 123–128.
24. He, X.; Lin, Z.S. Interactive Effects of the Influencing Factors on the Changes of $PM_{2.5}$ Concentration Based on GAM Model. *Environ. Sci.* **2017**, *38*, 22–32.
25. Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.A.; Rosaida, N.; Hamid, H.A. Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). *Atmos. Environ.* **2013**, *77*, 621–630. [CrossRef]
26. Abdul-Wahab, S.A.; Bakheit, C.S.; Al-Alawi, S.M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environ. Model. Softw.* **2005**, *20*, 1263–1271. [CrossRef]
27. Vaidya, O.C.; Howell, G.D.; Leger, D.A. Evaluation of the Distribution of Mercury in Lakes in Nova Scotia and Newfoundland. *Water Air Soil Pollut.* **2000**, *117*, 353–369. [CrossRef]

28. Debarchana, G.; Manson, S.M. Robust Principal Component Analysis and Geographically Weighted Regression Urbanization in the Twin Cities Metropolitan Area of Minnesota. *J. Urban Reg. Inf. Syst. Assoc.* **2008**, *20*, 15–25.
29. Zou, B.; Pu, Q.; Bilal, M.; Weng, Q.; Zhai, L.; Nichol, J.E. High-Resolution Satellite Mapping of Fine Particulates Based on Geographically Weighted Regression. *IEEE Geosci. Remote Sens.* **2016**, *13*, 495–499. [CrossRef]
30. Zou, B.; Xu, S.; Sternberg, T.; Fang, X. Effect of Land Use and Cover Change on Air Quality in Urban Sprawl. *Sustainability* **2016**, *8*, 677. [CrossRef]
31. An, F.; Zhai, L.; Sang, H.Y.; Zhang, Y.; Zhou, Y.; Yuan, J. Multiple regression analysis on $PM_{2.5}$ impact factors based on geographic conditions monitoring data. *Sci. Surv. Mapp.* **2015**, *40*, 58–63.
32. Meng, X.; Chen, L.; Cai, J.; Zou, B.; Wu, C.F.; Fu, Q.; Zhang, Y.; Liu, Y.; Kan, H. A land use regression model for estimating the NO_2 concentration in Shanghai, China. *Environ. Res.* **2015**, *137*, 308–315. [CrossRef] [PubMed]
33. Pearson, K. On lines and planes of closest fit to systems of points is space. *Philos. Mag. Ser.* **1901**, *62*, 559–572. [CrossRef]
34. Kabacoff, R.I. *R in Action: Data Analysis and Graphics with R*, 2nd ed.; Manning Publications Co.: Shelter Island, NY, USA, 2011; pp. 191–195.
35. Bae, J.; Kim, J.T.; Kim, J.H. Subset selection in multiple linear regression: An improved Tabu search. *J. Korean Soc. Mar. Eng.* **2016**, *40*, 138–145. [CrossRef]
36. Shi, N.; Cao, H.X. The Optimum Climate Forecasting Model Based on All Possible Rrgressions. *J. Nanjing Inst. Meteorol.* **1992**, *15*, 459–566.
37. Draper, N.R.; Smith, H. *Applied Regression Analysis*, 3th ed.; John Wiley & Sons: New York, NY, USA, 1998.
38. Ayón, P.; Swartzman, G.; Espinoza, P.; Bertrand, A. Long-term changes in zooplankton size distribution in the Peruvian Humboldt Current System: Conditions favouring sardine or anchovy. *Mar. Ecol. Prog. Ser.* **2011**, *422*, 211–222. [CrossRef]
39. Chen, C.C.; Wu, C.F.; Yu, H.L.; Chan, C.C.; Cheng, T.J. Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter $PM_{2.5}$ concentrations. *Atmos. Environ.* **2012**, *54*, 1–8. [CrossRef]
40. Liu, Y.; Paciorek, C.J.; Koutrakis, P. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* **2009**, *117*, 886–892. [CrossRef] [PubMed]
41. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [CrossRef] [PubMed]
42. Brown, D.G.; Goovaerts, P.; Bumlckl, A.; Li, M.Y. Stochastic Simulation of Land-Cover Change Using Geostatistics and Generalized Additive Models. *Photogramm. Eng. Remote Sens.* **2002**, *68*, 1051–1061.
43. Moreno-Torres, J.G.; Saez, J.A.; Herrera, F. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1304–1312. [CrossRef] [PubMed]
44. Chen, L.; Bai, Z.P.; Di, S.; You, Y.; Li, H.M.; Liu, Q. Application of land use regression to simulate ambient air PM_{10} and NO_2 concentration in Tianjin City. *China Environ. Sci.* **2009**, *29*, 685–691.