A Clustering Framework to Reveal the Structural Effect Mechanisms of Natural and Social Factors on PM$_{2.5}$ Concentrations in China

Wentao Yang 1, Zhanjun He 2, Huikun Huang 3 and Jincai Huang 4,*

1 National-Local Joint Engineering Laboratory of Geospatial Information Technology, Hunan University of Science and Technology, Xiangtan 411100, China; yangwentao8868@126.com
2 School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; hezj@cug.edu.cn
3 School of Geosciences and Info-physics, Central South University, Changsha, Hunan 410083, China; uanghuikun@csu.edu.cn
4 Shenzhen Key Laboratory of Spatial Smart Sensing and Service, Shenzhen University, Shenzhen 518060, China
* Correspondence: huangjincaiszu@szu.edu.cn; Tel.: +86-153-0849-4530

Abstract: Understanding the mechanisms of various factors that affect PM$_{2.5}$ can assist in the development of scientific measures to improve air quality. Nevertheless, existing research has concentrated on exploring local effect mechanisms, while structural effect mechanisms at regional or national scales have scarcely been analysed. Consequently, this study presents an analytical framework for elucidating the structural effect mechanisms of associated factors on PM$_{2.5}$. Geographically and temporally weighted regression was used to explore the local effect mechanisms. This was followed by spatial clustering analysis to reveal these mechanisms by detecting their aggregation pattern. In the analysis, datasets for annual mean PM$_{2.5}$ and socio-economic factors in China from 1999 to 2016 were employed. Urban population, gross industrial output, and sulphur dioxide emissions were identified as factors affecting changes in PM$_{2.5}$ concentrations. These three factors had both negative and positive effects, while the gross industrial output had the largest coefficient variation degree. Three geographically related factors exhibited different impacts on PM$_{2.5}$ concentrations in most of mainland China. These factors were the urban population roughly west of the Heihe-Tengchong line, gross industrial output primarily in southwestern China, and sulphur dioxide emissions primarily in southern China.

Keywords: air pollution; structural effect mechanisms; spatiotemporal variations; spatiotemporal regression models; spatial clustering analysis

1. Introduction

Access to clean air is a fundamental human right. However, many air pollution events have occurred globally in the past several decades and have had a serious impact on human physical and mental health [1–4]. According to a report by the World Health Organisation [5], air pollution has resulted in more than two million premature deaths each year since the beginning of the 21st century, especially in developing countries. Given this unprecedented situation, the United Nations 2030 agenda for sustainable development (published in 2015) clearly stated that countries should substantially reduce the number of illnesses and deaths from air pollution by 2030 [6]. This directive presents an unprecedented challenge for China and other countries with poor air quality.

To improve air quality, the mechanisms of factors associated with the air quality or pollutant concentrations must first be identified. Recent research has analysed the effect mechanisms of factors associated with PM$_{2.5}$ concentrations. With consideration of both
theory and methods, existing models for detecting effect mechanisms can be roughly divided into two categories: multivariate receptor models and statistical analysis models [7]. Multivariate receptor models, such as chemical mass balance [8] and positive matrix factorisation [9], are used to decompose PM$_{2.5}$ in terms of meteorology, physics, and chemistry. Then, the components and their sources can be identified. These models can simply and clearly reveal the source apportionment of PM$_{2.5}$ [10]. Nevertheless, they represent micro-scale analyses based on a natural science and ignore macro-scale socioeconomic factors.

In contrast, statistical analysis models are employed to describe the relationship between PM$_{2.5}$ concentrations and natural and/or socio-economic factors by means of random functions. These models can be further divided into two categories: statistical regression models and machine learning models. Both models can be applied to predict unknown PM$_{2.5}$ concentrations at given locations. However, machine learning models generally learn nonlinear relationships from data. Owing to an unknown mechanism or “black box” process, the parameters in machine learning models lack interpretability, causing the model to lack an inferential capability. In comparison, statistical regression models, such as land use regression (LUR) and geographically weighted regression (GWR), reveal the effects of factors associated with PM$_{2.5}$ by representing an explicit relationship between them [11–14]. Various studies based on these models have been conducted. The varying results significantly depended on the research area and time period. Especially in China, these differences are mainly manifested in two aspects.

First, the dominant factors affecting PM$_{2.5}$ concentrations may differ over space and time. For example, using the LUR model, Liu et al. [12] found that location factors, such as longitude and distance to the coast, were the most important factors in Shanghai, whereas traffic conditions were the most important in the Nanchang urban area [13]. When using GWR regression to analyse air quality with a $1 \text{ km} \times 1 \text{ km}$ grid across China from 2001 to 2010, Lin et al. [14] confirmed that the two main driving forces affecting PM$_{2.5}$ concentrations were local economic growth and urban expansion.

Second, the same factor may have different effects on PM$_{2.5}$ concentrations over space and time. For instance, in a study of the impact of the urbanisation level on urban air quality, Han et al. [15] found that significant positive correlations existed between urban populations and PM$_{2.5}$ concentrations. Similarly, through a study of 343 Chinese cities from 1998 to 2012, Luo et al. [16] confirmed that there was a very strong positive correlation between all socioeconomic factors and PM$_{2.5}$ concentrations. Wang and Fang [17] recognised that PM$_{2.5}$ concentrations showed a positive relationship with urbanisation rates, but a negative relationship with per capita gross domestic product (GDP) and the building industry in the Bohai Rim Urban Agglomeration, China. A 2014 empirical study of 190 Chinese cities undertaken by Zhou et al. [18] indicated that economic growth had a highly negative effect on PM$_{2.5}$ concentrations.

The above differences in factors and effects can be considered as spatiotemporal variations of the effects of factors associated with PM$_{2.5}$ concentrations. They are also referred to as spatiotemporal heterogeneity or non-stationary data in geographic information science when the statistical characteristics of univariate data or the relationships between multivariable changes over space and time are referenced [19]. According to differences in analysis units, spatiotemporal heterogeneity can be divided into local heterogeneity and structural/stratified heterogeneity [20]. Local heterogeneity emphasises spatiotemporal variations at the individual level, such as the effect mechanisms of factors associated with PM$_{2.5}$ concentration changes over a single city or other geographical analytical units. These are defined as local effect mechanisms. Because nearby cities or geographical analytical units often have similar geographical conditions and pollutant characteristics, local effect mechanisms are usually similar between cities or units, such that there are spatial aggregation structures among local effect mechanisms. Therefore, the structural heterogeneity of effect mechanisms indicates that some cities or multiple units with similar local
effect mechanisms can form a group. Accordingly, similar mechanisms are deemed structural effect mechanisms in related studies.

Regional pollution characteristics in China suggest that joint control of air pollution is more effective in improving air quality. Obtaining structural effect mechanisms can, therefore, provide important decision-making information for such efforts. However, most studies based on statistical analysis models focused on local heterogeneity and have not discussed the structural heterogeneity of the effect mechanisms. Therefore, this study aimed to uncover the structural effects of factors associated with PM$_{2.5}$ concentrations in China using spatiotemporal analysis.

2. Materials and Methods

2.1. Data

A total of 334 Chinese cities were selected as the basic analytical units (including county-level and prefectural-level cities, as well as municipalities) along with multisource datasets corresponding to these cities (Table 1).

Table 1. Multisource datasets used in this study.

| Category                          | Year       | Spatial Scale | Source                                      |
|-----------------------------------|------------|---------------|---------------------------------------------|
| Global Annual PM$_{2.5}$ Grid     | 1999–2016  | 10 km         | http://sedac.ciesin.columbia.edu/           |
| Night-time Satellite              | 1999–2016  | 1 km          | https://earthdata.nasa.gov/                 |
| Urban Population                  | 1999–2016  | 334 cites (China) | http://www.stats.gov.cn/                    |
| Gross Domestic Product            | 1999–2016  | 334 cites (China) | http://www.stats.gov.cn/                    |
| Gross Industrial Output           | 1999–2016  | 334 cites (China) | http://www.stats.gov.cn/                    |
| Urban Electricity Consumption     | 1999–2016  | 334 cites (China) | http://www.stats.gov.cn/                    |
| Sulphur dioxide emission          | 1999–2016  | 334 cites (China) | http://www.stats.gov.cn/                    |

Large-scale ground PM$_{2.5}$ monitoring in China was initiated in 2012. Global annual PM$_{2.5}$ grid data were selected to extract PM$_{2.5}$ concentrations in Chinese cities. This dataset was produced using the total columnal aerosol optical depth (AOD) obtained from a combination of moderate resolution imaging spectroradiometer, moderate resolution imaging spectroradiometer, sea-viewing wide-field-of-view sensor AOD satellite instruments, and coincident aerosol vertical profiles [21]. These data provide a series of three-year running average PM$_{2.5}$ concentrations with extensive coverage (from 70°N to 55°S), a long temporal range (from 1999 to 2016), and high resolution (~10 km). They have been widely used at regional and national scales. An alternative satellite-retrieved PM$_{2.5}$ dataset was acquired from Dalhousie University products from 1999 to 2017 [22]. To accurately obtain PM$_{2.5}$ concentrations in Chinese cities, DMSP/OLS (1999–2012) and NPP/VIIRS (2013–2016) night-time satellite datasets were collected to extract data of urban built-up areas (Figure 1). The statistical values of the PM$_{2.5}$ grids within the urban built-up areas were regarded as the PM$_{2.5}$ concentrations for each city.

The variable selection in this research was mainly based on previous studies and depended on whether complete data collection was possible. The socioeconomic data consisted of the urban population, GDP, gross industrial output, urban electricity consumption, and sulphur dioxide emissions (SO$_2$), all of which were based on cities as statistical units and could be directly used for analysis without data conversion. The spatial distribution of all variable values in 1999 is shown in Figure 2.
2.2. Methods

The local effect (the relationship between PM$_{2.5}$ and associated factors in each city) was obtained using spatiotemporally varying regression models whose parameters changed over space. Based on the first law of geography claiming that ‘everything is related to everything else, but near things are more related to each other’ [23], it was assumed that the local effect mechanisms between nearby cities were more similar and grew more different with increased distance between cities. Therefore, the whole study area could be divided into several sub-areas with similar relationships between PM$_{2.5}$ and the associated factors. Similar relationships in each sub-area were defined as structural effect mechanisms. Detecting these was essentially consistent with spatial clustering analysis, which aims to partition a dataset into several clusters according to spatial proximity and attribute similarity. Consequently, this approach was applied to detect structural effect mechanisms by analysing the aggregation pattern of local effect mechanisms.
The analytical process used in this study consisted of two steps (Figure 3). First, geographically and temporally weighted regression (GTWR) was used to identify the local effect mechanisms of natural and socio-economic factors on PM$_{2.5}$ concentrations [24]. Comparison models, such as GWR and ordinary least squares regression (OLS), were employed to evaluate the effectiveness of the regression results of the GTWR model. Second, a spatial clustering method, the regionalisation with dynamically constrained agglomerative clustering and partitioning algorithm (REDCAP), was used to determine the structural effect mechanisms of different associated factors on PM$_{2.5}$ concentrations [25]. Different types of clustering validity evaluation indices were then applied to determine the optimal clustering results.

Figure 3. Main analytical process used in this study.

2.2.1. Identifying Local Effect Mechanisms Using the GTWR Model

Spatial heterogeneity means that the relationships between the input variables and output variables are not constant in the given area, and a global model cannot reveal the spatial variation in the relationships among the spatial data [26]. GWR is one of the most widely used localised models for dealing with spatial heterogeneity [27]. It enables exploration of local effect mechanisms and testing of variation significance, thereby fostering research attention on the atmospheric environment [11]. Despite the success of GWR in addressing spatial variations, it remains a challenge when temporal heterogeneity is present in dynamic geographical data. For dynamic geographical data, the relationships among different variables may not only change over space but also vary at different timestamps, namely the existence of spatial and temporal heterogeneity. Hence, the GTWR model aims to build a series of local models whose parameters vary across space–time locations to handle issues involving both spatial and temporal heterogeneity [24].

Assuming an observation sample is labelled $i$ ($i = 1, 2, ..., n$), where $n$ is the number of observations, the GTWR model can be mathematically described as:

$$PM_{2.5}(u_i, v_i, t_i) = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i)X_{ik} + \epsilon(u_i, v_i, t_i),$$

where $(u_i, v_i, t_i)$ denotes the location of the observation sample $i$ in space $(u_i, v_i)$ at time $t_i$, $PM_{2.5}(u_i, v_i, t_i)$ and $X_{ik}$ are the PM$_{2.5}$ concentration value and relevant factor values ($k$ is the total number of relevant factors), respectively, $\beta_0(u_i, v_i, t_i)$ denotes the intercept term, $\beta_k(u_i, v_i, t_i)$ represents the slope parameter of the $k^{th}$ factors, which describe the relationship between the $k^{th}$ factors and PM$_{2.5}$, and $\epsilon(u_i, v_i, t_i)$ is the error term. The model allows the parameters to vary across space and time and can, thus, capture the local effect of these dimensions.

For calibration, the GTWR model assumes that observation data close to sample $i$ have a greater influence on the estimation of the parameters $\beta(u_i, v_i, t_i)$ than observation data located further away. Therefore, the estimation of the parameters $\beta_k(u_i, v_i, t_i)$ is expressed as:
\begin{equation}
\hat{\beta}(u_i, v_i, t_i) = [X^T W(u_i, v_i, t_i) X]^{-1} X^T W(u_i, v_i, t_i) Y,
\end{equation}

where the weight matrix $W(u_i, v_i, t_i)$ is equal to $\text{diag}(w_{i,1}, w_{i,2}, ..., w_{i,n})$, and the weight parameter $w_{i,j}$ ($j = 1, 2, ..., n$) represents the contribution of observation sample $j$ to the estimation of parameters $\beta(u_i, v_i, t_i)$. The estimation of the weight parameter $w_{i,j}$ depends on the spatiotemporal distance function between observation samples $j$ and $i$. In general, a Gaussian decay-based function is used to determine the weight value:

\begin{equation}
w_{i,j} = \exp \left( \frac{(d_{ij}^ST)^2}{(h_{ST})^2} \right),
\end{equation}

where $d_{ij}^ST$ denotes the spatiotemporal distance between observation samples $j$ and $i$, and $h_{ST}$ is the spatiotemporal bandwidth parameter that is employed to produce a decay of influence with respect to distance. Considering the different scale effects of space and time, an ellipsoidal coordinate system is applied to measure the spatiotemporal distance $d_{ij}$. Two types of parameters should be determined in the GTWR model: one is the regression coefficient parameter; the other is the spatiotemporal bandwidth parameter. The regression coefficients can be directly obtained using Equation (2). However, the weight matrix in that equation is dependent on the spatiotemporal bandwidth parameter shown in Equation (3). It should, thus, be determined first. Therefore, the regression coefficients are different under different spatiotemporal bandwidth parameters. The optimal parameters and the corresponding regression coefficients can be determined by evaluating the Akaike information criterion (AIC) or cross-validation function value under different spatiotemporal bandwidth parameters [27,28]. All the work done in this study used the ArcGIS 10.2 platform with the GTWR_Beta package [24].

### 2.2.2. Detecting Structural Effect Mechanisms Using the REDCAP Algorithm

Although a variety of spatial clustering algorithms have been developed to detect aggregation patterns, a hierarchical method known as REDCAP was adopted in this study. It guarantees spatial proximity and attribute similarity within clusters or areas [25,29]. The REDCAP algorithm consists of two steps: a hierarchical clustering strategy, which generates a spatially contiguous tree, and average linkage, which defines the similarity of two clusters. The spatially contiguous tree is then partitioned into several subtrees by optimising an objective function, such as the sum of squared differences (SSD), which is defined as:

\begin{equation}
\text{SSD} = \sum_{k=1}^{l} \sum_{p=1}^{m_k} \sum_{q=1}^{n} (\hat{\beta}(u_p, v_p, t_q) - \bar{\beta}(k, t_q))^2,
\end{equation}

where $l$ is the number of clusters, $m_k$ denotes the number of data objects (cities in this study) in cluster $k$, $n$ represents the number of variables considered (time length in this study), $\hat{\beta}_{pq}$ represents a variable value (estimated value of the regression parameters for the $p$th city in cluster $k$ at time), and $\bar{\beta}(k, t_q)$ represents the mean value of the regression parameter in cluster $k$ at time $t_q$.

Figure 4 shows the clustering process for a simulation dataset using the REDCAP method. In Figure 4a, the shade of each point (here regarded as a city) represents the effect value of a certain factor (such as urban population) on PM2.5 concentrations. The dotted lines are the edges of the Delaunay triangulation network, which is generated based on the spatial locations of all points (cities). Each edge is linked to an attribution value to describe the similar degrees between the effect values of these two points (cities). The minimum spanning tree (the spatially contiguous tree) shown in Figure 4b can be constructed based on the connected network. Next, the tree is partitioned into several subtrees. First, the tree is partitioned by removing edge (4, 5) to create two structural regions/clusters, as shown in Figure 4c, based on the SSD measurement. Next, the best cut (9, 10) is found in
the remaining edges to further create three structural regions/clusters (Figure 4d), and the above steps are repeated to generate more structural regions/clusters.

![Figure 4. Illustration of the regionalisation with dynamically constrained agglomerative clustering and partitioning (REDCAP) algorithm clustering process.](image)

- **a.** each point represents the effect value of a certain factor on PM$_{2.5}$ concentrations;
- **b.** the minimum spanning tree can be constructed based on the connected network;
- **c.** tree is partitioned by removing edge (4,5) to create two structural regions/clusters;
- **d.** the best cut (9,10) is found in the remaining edges to further create three structural regions/clusters and the above steps are repeated to generate more structural regions/clusters.

Use of the REDCAP algorithm enables the partitioning results with different numbers of clusters or regions to be obtained. Clustering evaluation is a key task to determine whether a high-quality clustering result is obtained. Existing clustering evaluation indices can be divided into two categories: external evaluation and internal evaluation [30–32]. The first category is statistically complex and requires prior knowledge of the clustering results, whereas the second category does not. The best clustering result is always obtained by comparing the results of different algorithms or different clustering parameters. In this research, we had no prior knowledge of the PM$_{2.5}$ data. Two representative internal evaluation indices, Silhouette (Sil) and Davies–Bouldin (DB) indices, were selected to identify the validity of the partitioning result [33,34]. A better clustering result generally corresponds to a larger Sil value or a smaller DB value. The change curves of the Sil and DB indices thus enable the optimal partition result to be obtained.

3. Results

3.1. Identification of Local Effect Mechanisms of Associated Factors on PM$_{2.5}$ Concentrations

The GTWR model was first applied to reveal the spatiotemporal variation of the relationships between PM$_{2.5}$ and other variables. To evaluate the effectiveness of the results, four accuracy evaluation indices were used to analyse the performance of the GTWR, OLS, and GWR models. The AIC value of the GTWR model (45179) was smaller than those of the OLS (53321) and GWR (50341) models (Table 2). The $R^2$ and $R^2_{adj}$ values of the GTWR model were larger than those of the OLS and GWR models. Out-of-sample validation was used to further evaluate the model performance (80% random samples for modelling and the remaining 20% for validation/prediction analysis). The GTWR prediction results had the smallest root mean square error (RMSE) value (12.47), followed by GWR (18.75) and OLS (25.81). All index values indicated that the GTWR model outperformed the OLS and GWR models when describing spatiotemporal variations in PM$_{2.5}$ concentrations and other variables. Using the variance inflation factor to detect the amount of multi-collinearity and a t-test to determine the statistical significance of a regression coefficient, three variables were identified as being the most closely associated with changes in PM$_{2.5}$ concentrations in the study area: urban population (unit: 1 million), gross industrial output (unit: ¥ 10 billion), and sulphur dioxide emissions (unit: 10,000 t).

### Table 2. Accuracy evaluation results for different models.

| Index | GTWR | OLS  | GWR  |
|-------|------|------|------|
| AICc  | 45,179 | 53,321 | 50,341 |
The spatial distributions of the regression coefficients for the different variables in 2004, 2010, and 2016 (Figures 5–7) showed that they had an obvious spatial aggregation structure; for example, the estimated coefficient values at nearby locations tended to be closer. It is worth noting that, to clearly show the spatial distribution of the regression coefficients in Figures 5 to 7, the estimated values of these regression coefficients for all cities were labelled for the whole administrative region. The minimum values of the regression coefficients of these three variables were negative (−0.76, −0.59, and −0.26; Table 3), whereas the maximum values (9.1, 2.41, and 1.32) were positive. This finding shows that all variables had both negative and positive effects on PM$_{2.5}$ concentrations.

According to the mean values of the regression coefficients, when the population, gross industrial output, and sulphur dioxide emissions increased by 1 million, ¥ 10 billion, and 10,000 t, respectively, the PM$_{2.5}$ concentration increased by approximately 1.54, 0.05, and 0.09 ug/m$^3$ on average, respectively. These three associated factors always exhibited positive effects on PM$_{2.5}$ concentrations. Additionally, the coefficient of variance (CV), which was used to measure the degree of dispersion of variables with different units, was calculated by the ratio of the standard deviation (Std.) to the mean value. Gross industrial output had the largest coefficient variation, followed by sulphur dioxide emission and urban population.

Table 3. Descriptive statistics of regression coefficients for different variables.

| Variable                        | Descriptive Statistics of Regression Coefficients |
|---------------------------------|--------------------------------------------------|
|                                 | Min     | Max     | Mean   | Std.   | CV   |
| Urban population (1 million)    | −0.76   | 9.10    | 1.54   | 1.72   | 1.12 |
| Gross industrial output (¥ 10 billion) | −0.59   | 2.41    | 0.05   | 0.28   | 5.60 |
| Sulphur dioxide emission (10,000 t) | −0.26   | 1.32    | 0.09   | 0.24   | 2.67 |

Figure 5. Spatial distribution of the regression coefficient for the population in 2004, 2010, and 2016.

Figure 6. Spatial distribution of the regression coefficient for the gross industrial output in 2004, 2010, and 2016.
3.2. Identification of the Structural Effect Mechanisms of Associated Factors on PM$_{2.5}$ Concentrations

The regression coefficients exhibited obvious spatial aggregation. Therefore, the REDCAP algorithm was used to quantitatively identify the aggregation structure for each one. Based on the similarity of regression coefficients of the population, all cities were divided into 2 to 10 clusters. The Sil and DB indices were then applied to determine the optimal clustering results. Although neither polyline (Figure 8a,b) had extrema values (a maximum Sil value and a minimum DB value) in the intervals, it is reasonable that the boundary point, corresponding to a relatively large Sil value and a relatively small DB value [35], was identified as the optimal parameters of the clustering structures. The cluster number should be two when the Sil value is relatively large and the DB value is relatively small. The spatial distribution of the obtained two clusters is shown in Figure 9, and the corresponding statistical result is shown in Figure 8c.

Cluster I mainly included most prefecture-level cities in northwestern China, while the remaining areas composed Cluster II. The mean coefficient values in Cluster I were approximately 5 μg/(m$^3$·million) before 2008 and gradually decreased to 3 μg/(m$^3$·million) in 2016. However, in Cluster II, these values were 1 μg/(m$^3$·million) during the entire study period. These results indicated that population had a greater impact on PM$_{2.5}$ concentrations in Cluster I than in Cluster II. These clusters were roughly divided by the Heihe-Tengchong line, an imaginary boundary dividing China diagonally into two parts: the area east of the line contained 43% of the country’s land and 94% of the total population, while the area west of the line contained 57% of the land but only 6% of the population. PM$_{2.5}$ had lower values west of this line (i.e., Cluster I), where the main sources were sand and dust, and higher values east of this line (i.e., most of Cluster II), where the main sources were emissions from human activity [36]. Because of the high level of PM$_{2.5}$ caused by human activity in Cluster I, PM$_{2.5}$ change may be less sensitive to changes in population than in Cluster II. Therefore, the regression coefficients in Cluster II were lower than those in Cluster I. In addition, 2006 was the first year of the 11th Five-Year Plan of China, in which building a resource-saving and environment-friendly society was first added. The implementation of effective measures for controlling pollutants meant that PM$_{2.5}$ in most cities showed a decreasing trend from 2006 onward. Consequently, the relatively high coefficient of Cluster I meant that the regression coefficient was impacted severely and exhibited a decreasing trend.
Cluster validity was evaluated based on the regression coefficient series of the gross industrial output (Figure 10a,b). When the cluster number was equal to 2, the Sil and DB indices corresponded to the largest and smallest values, respectively. Therefore, the study area was divided into two sub-areas based on the regression coefficients of the population. Cluster I mainly included cities in Sichuan Province, Guizhou Province, and Chongqing Municipality (Figure 11). The coefficient values first decreased and then stabilised after 2006. In contrast, the temporal variation in the coefficient values of gross industrial output in Cluster II was lower than in Cluster I, and the average values from 1999 to 2016 fluctuated around $0 \mu g/(m^2\cdot¥ 10$ billion). Before 2006, the coefficient values in Cluster I were higher than those in Cluster II; thereafter, both values were equal. According to statistical data on industrialisation levels in China, southwestern China lagged behind other areas before 2000.

Owing to development in western China since 2000, industrial development in southwestern China has achieved remarkable progress. Meanwhile, implementing the ‘new industrialisation’ strategic plan in China has caused a series of severe problems in this region, including ecological environment destruction caused by rapid industrial development. This issue has received greater attention since 2000 [37,38]. It is possible that the optimisation of the industrial structure and the improvement of industrialisation quality will lead to a gradual reduction in the impact of the change in gross industrial output on air pollution in southwestern China. Since 2005, the overall impact in the southwest...
was close to those of other cities in China, resulting in the equal regression coefficients shown in Figure 10.

![Figure 10. Clustering of gross industrial output regression coefficients.](image)

The cluster validity evaluation of the sulphur dioxide emission regression coefficient (Figure 12) showed that when the cluster number was set as 2, the Sil and DB indices corresponded to the largest and smallest values, respectively. Hence, it was reasonable to divide the study area into two clusters: Cluster I was mainly located in southern China, including Guangdong Province and parts of Fujian and Hubei Provinces, with other areas belonging to Cluster II (Figure 13). Both clusters had similar coefficient value change curve morphologies: values first remained steady and then increased gradually after 2006. However, the degree of increase in Cluster II was larger than that in Cluster I. Strong observational evidence has indicated that aerosols in southern China are mainly composed of sulphate derived from chemical reactions between sulphur dioxide and other substances [39]. After 2006, Cluster I experienced a decreasing trend in the time series of sulphur dioxide emission, suggesting that the change in sulphur dioxide emission had a greater influence on PM$_{2.5}$ concentrations in southern China than in the other areas.

![Figure 11. Spatial distribution of different clusters of gross industrial output.](image)
4. Conclusions

This study analysed spatiotemporal variations in the effect mechanisms of associated factors on PM$_{2.5}$ in China from 1999 to 2016. The effect mechanisms were divided into two categories: local (analysing spatiotemporal variations in each region at the individual unit/city level) and structural (extracting aggregation patterns of multiple units from the group level). The GTWR model was used to explore the local effect mechanism by modelling the relationships between PM$_{2.5}$ concentrations and associated factors. Three variables—urban population, gross industrial output, and sulphur dioxide emissions—were identified as being the most closely associated with changes in PM$_{2.5}$ concentrations. All variables had both negative and positive effects on PM$_{2.5}$ concentrations, while gross industrial output and urban population had the largest and smallest degrees of coefficient variation, respectively. The REDCAP algorithm was used to detect the structural effect mechanism by dividing the study area into several quasi-homogeneous sub-areas based on similarities in the change curves of the regression coefficients. Two clusters (or spatial sub-areas) were identified for these three variables.

Regarding the spatiotemporal variation of the effect mechanisms of socio-economic factors, previous studies have shown that PM$_{2.5}$ pollution is greater in more populated cities on account of daily living and production activities. Their correlation to polluting gas emissions and higher population levels always lead to greater energy consumption and increased emissions [36]. However, the relationships between population and gas emissions are not constant over space and time. For example, vehicle emissions were re-
Sustainability, 2021, 13, 1428

Regarding as one of the major sources of PM$_{2.5}$ pollution in China [41]. Owing to the spatio-temporal difference of consumption levels and habits, the same population increase may cause an increase in different vehicle usage, which leads to the different change of PM$_{2.5}$ concentrations in different geographical areas and time periods.

Similarly, the use of fossil fuels increases with the industrial development in a region [42], which inevitably increases the emission of atmospheric pollutants. Nevertheless, the structure of industry changes across different cities and times; hence, the same increases of gross industrial output may result in a different change of PM$_{2.5}$ in different areas and times. Anthropogenic emission of sulphur dioxide plays a critical role in the process of secondary fine particulate matter formation [43]. The secondary pollution depends on the related factors, such as weather and other geographical factors. The spatiotemporal heterogeneity of these factors will cause sulphur dioxide emissions to have different geographical effects on PM$_{2.5}$. In addition, Figures 8c, 10c, and 12c show that 2006 was the approximate turning point for the effect mechanisms of socioeconomic impacts on PM$_{2.5}$ in many areas. Socioeconomic impacts are always related to national and local policies. The year 2006 was the first of the 11th Five-Year Plan of China, in which building a resource-saving and environment-friendly society was first added. A series of pollution control measures have since been implemented to change the effect mechanisms of socioeconomic factors on PM$_{2.5}$ from 2006 onward.

Despite the contributions of this study, it has some limitations that need future consideration. First, although the GTWR model was effective at modelling the local effect mechanism, it could not describe the possible nonlinear relationships between PM$_{2.5}$ concentrations and the associated factors. Therefore, a method for simultaneously modelling linearity and spatiotemporal patterns must be studied. Second, this study concentrated on describing spatiotemporal variations of the effect mechanisms and only briefly discussed the possible reasons. Quantitatively exploring the political functions of the effect mechanisms should be a subject for further research.

Author Contributions: Conceptualisation, W.Y. and Z.H.; methodology, W.Y. and H.H.; software, J.H.; validation, W.Y. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the funds of the National Science Foundation of China (grant number 41801311), the Philosophy and Social Science Foundation of Hunan Province, China (grant number 18YBQ050), and the Scientific Research Fund of Hunan Provincial Education Department (grant number 19C0777).

Institutional Review Board Statement: Ethical review and approval were waived because necessary permissions were obtained from the local governments to which the schools are affiliated for this study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request due to ethical restrictions. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to not having participants consent to share their anonymized data with a third party.

Acknowledgments: We express our sincere appreciation to the anonymous reviewer for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.
References

1. Schwarzenbach, R.P.; Egli, T.; Hofstetter, T.B.; Von Gunten, U.; Wehrli, B. Global Water Pollution and Human Health. *Annu. Rev. Environ. Resour.* 2010, 35, 109–136, doi:10.1146/annurev-environ-100809-125342.

2. Landrigan, P.J.; Fuller, R.; Hu, H.; Caravano, J.; Cropper, M.L.; Hanrahan, D.; Sandilya, K.; Chiles, T.C.; Kumar, P.; Suk, W.A. Pollution and global health—An agenda for prevention. *Environ. Health Perspect.* 2018, 126, 084501.

3. Tu, J.; Tu, W. How the relationships between preterm birth and ambient air pollution vary over space: A case study in Georgia, USA using geographically weighted logistic regression. *Appl. Geogr.* 2018, 92, 31–40, doi:10.1016/j.apgeog.2018.01.007.

4. Hoek, G. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr. Environ. Health Rep.* 2017, 4, 450–462, doi:10.1007/s40572-017-0169-5.

5. World Health Organization. *WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide: Global Update 2005: Summary of Risk Assessment*. No. WHO/SDE/PAHO/06.02; World Health Organization: Geneva, Switzerland, 2006.

6. Colglazier, W. Sustainable development agenda: 2030. *Science* 2015, 349, 1048–1050, doi:10.1126/science.aad2333.

7. Cheng, Z.; Li, L.; Liu, J. Identifying the spatial effects and driving factors of urban PM2.5, pollution in China. *Ecol. Indic.* 2017, 82, 61–75, doi:10.1016/j.ecolind.2017.06.043.

8. Villalobos, A.M.; Barraza, F.; Jorquera, H.; Schauer, J.J. Chemical speciation and source apportionment of fine particulate matter in Santiago, Chile. 2013. *Sci. Total Environ.* 2015, 512, 133–142, doi:10.1016/j.scitotenv.2015.01.006.

9. Milano, C.; Huang, L.; Batterman, S. Trends in PM2.5 emissions, concentrations and apportionments in Detroit and Chicago. *Atmos. Environ.* 2016, 129, 197–209, doi:10.1016/j.atmosenv.2016.01.012.

10. Zikova, N.; Wang, Y.G.; Yang, F.M.; Li, X.H.; Tian, M.; Hopke, P.K. On the source contribution to Beijing PM2.5 concentrations. *Atmos. Environ.* 2016, 134, 84–95, doi:10.1016/j.atmosenv.2016.03.047.

11. Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Guilliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* 2008, 2, 7561–7578.

12. Liu, C.; Henderson, B.H.; Wang, D.; Yang, X.; Peng, Z.R. A land use regression application into assessing spatial variation of intra-urban fine particulate matter (PM2.5) and nitrogen dioxide (NO2) concentrations in City of Shanghai, China. *Sci. Total Environ.* 2016, 565, 607–615, doi:10.1016/j.scitotenv.2016.03.189.

13. Yang, H.; Chen, W.; Liang, Z. Impact of Land Use on PM2.5 Pollution in a Representative City of Middle China. *Int. J. Environ. Res. Public Health* 2017, 14, 462, doi:10.3390/ijerph14050462.

14. Lin, G.; Fu, J.; Jiang, D.; Hu, W.; Dong, D.; Huang, Y.; Zhao, M.D. Spatio-Temporal Variation of PM2.5 Concentrations and Their Relationship with Geographic and Socioeconomic Factors in China. *Int. J. Environ. Res. Public Health* 2014, 11, 173–186, doi:10.3390/ijerph110100173.

15. Han, L.; Zhou, W.; Li, W.; Li, L. Impact of urbanization level on urban air quality: A case of fine particles (PM 2.5) in Chinese cities. *Environ. Pollut.* 2014, 194, 163–170, doi:10.1016/j.envpol.2014.07.022.

16. Luo, J.; Du, P.; Samat, A.; Xia, J.; Che, M.; Xue, Z. Spatiotemporal Pattern of PM2.5 Concentrations in Mainland China and Analysis of Its Influencing Factors using Geographically Weighted Regression. *Sci. Rep.* 2017, 7, 40607, doi:10.1038/srep40607.

17. Wang, Z.B.; Fang, C.L. Geographically Weighted Regressions: A Method for Exploring Spatial Non-stationarity. *Geogr. Anal.* 1996, 28, 281–298, doi:10.1111/j.1538-4632.1996.tb00936.x.
28. Chen, J.; Zhou, C.; Wang, S.; Hu, J. Identifying the socio-economic determinants of population exposure to particulate matter (PM2.5) in China using geographically weighted regression modeling. Environ. Pollut. 2018, 241, 494–503, doi:10.1016/j.envpol.2018.05.083.

29. Deng, M.; Liu, Q.L.; Wang, J.Q.; Shi, Y. A general method of spatio-temporal clustering analysis. Sci. China Inf. Sci. 2013, 56, 1–14, doi:10.1007/s11432-011-4391-8.

30. Rand, W.M. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 1971, 66, 846–850.

31. Rendón, E.; Abundez, I.M.; Gutierrez, C.; Zagal, S.D.; Arizmendi, A.; Quiroz, E.M.; Arzate, H.E. A comparison of internal and external cluster validation indexes. In Proceedings of the 2011 American Conference, San Francisco, CA, USA, 2011, 29, pp. 1–10.

32. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. J. Intell. Inf. Syst. 2001, 17, 107–145, doi:10.1023/a:1012801612483.

33. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. 1979, 1, 224–227, doi:10.1109/tpami.1979.4766009.

34. Roussseuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 1987, 20, 53–65, doi:10.1016/0377-0427(87)90125-7.

35. Kryszczuk, K.; Hurley, P. Estimation of the number of clusters using multiple clustering validity indices. In Multiple Classifier Systems; Springer: Berlin/Heidelberg, Germany, 2010; pp. 114–123.

36. Qian, J.L.; Liu, C.S. Distributions and changes of aerosol optical depth on both sides of HU Huangyong Line and the response to Land use and land cover. Acta Sci. Circum. 2018, 38, 752–760.

37. Yin, C.B. Environmental efficiency and its determinants in the development of China’s western regions in 2000–2014. Chin. J. Popul. Resour. Environ. 2017, 15, 157–166, doi:10.1080/10042857.2017.1327687.

38. Liu, C.R.; Liu, H.Y.; Li, Y.F.; Li, Y.L. Socioeconomic Drivers of PM2.5 in the Accumulation Phase of Air Pollution Episodes in the Yangtze River Delta of China. Int. J. Environ. Res. Public Health 2016, 13, 928, doi:10.3390/ijerph13100928.

39. Zhao, S.; Xu, Y. Exploring the Spatial Variation Characteristics and Influencing Factors of PM2.5 Pollution in China: Evidence from 289 Chinese Cities. Sustainability 2019, 11, 4751, doi:10.3390/su11174751.

40. Behera, S.N.; Sharma, M. Degradation of SO2, NO2 and NH3 leading to formation of secondary inorganic aerosols: An environmental chamber study. Atmos. Environ. 2011, 45, 4015–4024.