DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.4, NOV. 2022

Spatial agglomeration and influencing factors of pulmonary tuberculosis in the Chinese mainland from 2015 to 2019

Jinguo Xin[a], Shuaixi Ma[b], Honghong Zhang[b]

a. Center of Information and Economy Social Development, Hangzhou Dianzi University, Hangzhou, China

b. College of Economics, Hangzhou Dianzi University, Hangzhou, China

ABSTRACT

Pulmonary tuberculosis is an infectious disease caused by Mycobacterium tuberculosis, which currently has a high incidence worldwide. Based on the incidence and mortality data of tuberculosis in 31 provinces in Chinese mainland from 2015 to 2019, this paper analyzes and studies the spatial agglomeration and local accumulation effects of tuberculosis incidence in China. Studies show that there is an obvious seasonal trend in the transmission of tuberculosis. The incidence peaks in March each year after the bottom in February. In addition, the tuberculosis incidence shows obvious spatial agglomeration, with a relatively high incidence rate in Xinjiang, Tibet, Qinghai, Guizhou, etc., while a low incidence rate in Beijing, Tianjin, Shanghai, Shandong, Jiangsu, Zhejiang, etc. Overall, the incidence has a clear upward trend from east to west, and the incidence rate in inland areas is higher than that in coastal areas. In addition, the paper also considers other factors that contribute to high-frequency transmissions, such as bad climate, poor air quality, and an underdeveloped economy. The most influential factor to the tuberculosis incidence is economic development level, and poor air quality in the northwest and southwest regions is also important reason for the high tuberculosis incidence.

KEYWORDS

Pulmonary tuberculosis; Spatial agglomeration effect; Cluster analysis

1 Introduction

At the end of 2019, an unexpected disaster swept across the world. Within less than three months, more than 70000 people in China were diagnosed with novel coronavirus pneumonia, panicking everyone for a time. People would feel nervous and helpless when they encounter unknown enemies since they are vulnerable. Tuberculosis was once called the "white plague", the mortality of which was so high that had caused great panic among human beings (Hussain, 2020). There are some differences as well as similarities between Tuberculosis and COVID-19. Tuberculosis is a chronic infectious disease caused by Mycobacterium tuberculosis, which can invade many organs, with pulmonary tuberculosis infection being the most common (Kumar et al., 2020). After the human body is infected with tuberculosis bacteria, the patient does not necessarily reveal symptoms (Talpur et al., 2020). Only when the

resistance is dampened or the cell-mediated allergy increases, it may cause clinical disease. When people have close contact with tuberculosis, the onset can be either rapid or slow, mostly low-grade fever (mainly in the afternoon), night sweats, fatigue, anorexia, weight loss, female menstrual disorders, etc., respiratory symptoms including cough, sputum, hemoptysis, chest pain, chest distress or breathing difficulty to different degrees (Peng et al., 2020; Talpur et al., 2020). It is spread mainly through the respiratory tract and digestive tract (Dimala et al., 2020; César et al., 2021). Generally speaking, the following groups of people are more susceptible to tuberculosis: 1. The malnourished people, especially those who live in crowded environments; 2. Infants with an immature cellular immune system; 3. The elderly, especially those with chronic diseases, such as hypertension, diabetes, coronary heart disease, chronic obstructive pulmonary disease, etc; 4. HIV-infected people; 5. Immunosuppressant users (Kadia et al., 2020; Ejemot-Nwadiaro et al., 2020).

China faces the severe challenge of tuberculosis, with a great number of domestic carriers, the infected, and deaths. China is ranked among one of the 22 countries with a heavy burden of tuberculosis in the world. In 2019, the number of pulmonary tuberculosis cases in China totaled 775764, with a 2990 death toll. The incidence rate and mortality rate were 55.5491/100000 and 0.2141/100000, respectively. The number of tuberculosis cases is still relatively high, and the prevention and treatment of tuberculosis in the central and western regions and rural areas remains tough. Therefore, the study on the areas and outbreaks of tuberculosis in China could provide a scientific basis for the rational prevention and control of its spread.

2 Material and Methods

The morbidity and mortality data mentioned in this paper covering 31 provinces in China from 2015 to 2019 (excluding Taiwan, Hong Kong and Macau) are excerpted from the China Health and Health Statistics Yearbook. The yearbook is an informative annual report reflecting the development of health services and the health status of residents in China. This book contains statistical data on the development of health services and the current health status of residents in 31 provinces, autonomous regions, and municipalities in Chinese mainland, as well as national statistical data in important historical years. The maps of China can be downloaded from the standard map service website of the Ministry of Natural Resources.

2.1 Spatial autocorrelation analysis

Spatial trend analysis: Based on the incidence and mortality data of tuberculosis in China from 2015 to 2019, a map of the spatial distribution of tuberculosis incidence rates from 2015 to 2019 was drawn to analyze the characteristics of the spatial distribution of tuberculosis in China.

Global spatial autocorrelation analysis: the Global Moran's I index is used in this paper to judge whether there was global spatial autocorrelation among 31 provinces. The Global Moran's I index is between -1 and 1: If I is more than 0 and value P is less than 0.05, there is a significant correlation between research variables in an adjacent region: if the Global Moran's I index is equal to or close to 0, there is no spatial self-correlation, which means that the data is randomly distributed (Getis & Ord, 2010; Liu et al., 2018). If the Global Moran's I index is less than 0 and the value P is less than 0.05, it indicates that there is a negative spatial correlation between research variables in the adjacent region (Anselin, 1995; Ord & Getis, 2010). The Global Moran's I index calculation formula is as follows:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where n represents the number of spatial regions studied. w_{ij} is the spatial weight of region i and j, reflecting the spatial relationship between region i and j. If region i is adjacent to region j, w_{ij} =1; otherwise w_{ij} =0. x_i and x_j are research variables (such as the incidence) in the region i and j, respectively, and x is the average value of the research variable (such as the average incidence).

2.2 K–Means clustering

K-means clustering is one of the simplest and most commonly used clustering algorithm. k cluster centers should be initialized first {C_1 , C_2 , C_3 … C_k $1< k \leq n$}. Then the Euclidean distance from each object to each cluster center is calculated (Zhao & Zhou, 2021; Mai et al., 2019), as shown in the following equation:

$$dis(x_i, c_j) = \sqrt{\sum_{t=1}^{m}(x_{it} - c_{jt})^2}$$

Where x_i is the ith object, c_j is the jth cluster center, x_{it} is the tth attribute of the ith object, c_{jt} is the tth attribute of the jth cluster center.

In each subsequent iteration, each object is reassigned to the nearest cluster based on the distance of each remaining object in the data set from the center of each cluster, until the algorithm converges or reaches the maximum number of iterations.

2.3 Hierarchical clustering

The hierarchical clustering method is a common clustering method used at home and abroad, and its basic idea is: The nearest sample is clustered into a class first, and the distant sample is clustered into a class later, likewise, each sample can be eventually clustered into the appropriate class (Šulc & Řezanková , 2019; Gregorius, 2004). The similarity of samples is measured with sample interval. There are many methods to calculate sample distances and euclidean distances is chosen in this paper. The binary Euclidean distance is shown below:

$$d(x_i, x_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$$

$d_{ij} = d(x_i, x_j), D = (d_{ij})_{p\times p}$, form a distance matrix:

$$\begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{2n} & \ldots & 0 \end{bmatrix}$$

Where $d_{ij} = d_{ji}$, the distance between the variables i and j.

The two nearest samples are combined into a class in the distance matrix, and after C_p merging with C_q into C_r, the distance from the other C_k is:

$$D_{r,k}^2 = \frac{n_k + n_p}{n_r + n_k} D_{pk}^2 + \frac{n_k + n_q}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

where n_p, n_k, n_r, n_q are the number of samples in C_p, C_k, C_r, C_q, respectively.

3 Result

3.1 Spatial distribution characteristics of tuberculosis incidence

Figure 1 shows that the incidence and mortality of tuberculosis reported in China have fluctuated in the past five years. The tuberculosis incidence has been declining year by year from 2015 to 2019, revealing an increasing level of tuberculosis prevention in China. On the other hand, the death rate peaked in 2018 before a slight drop in 2019. In addition, the number of reported tuberculosis cases each month in China shows an obvious seasonal trend. Most cases occur at the beginning of each year and gradually decrease in summer, autumn and winter. This is inevitably associated with the transmission mode of tuberculosis.

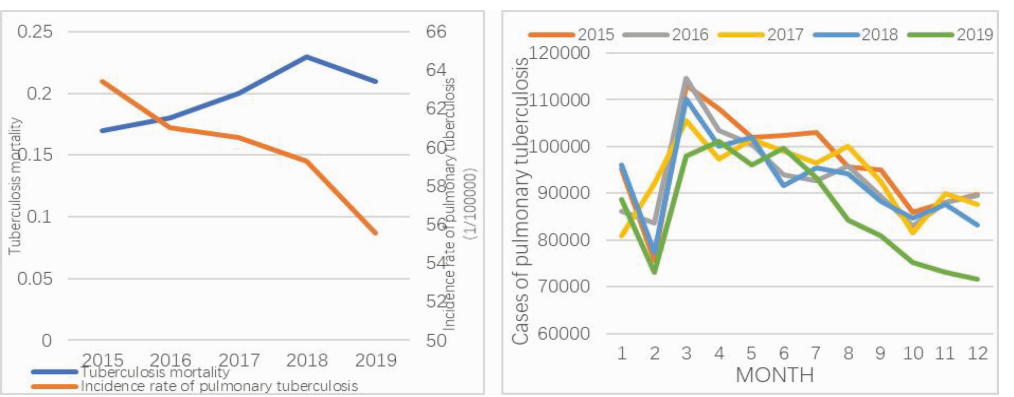


Figure 1 Seasonal trend chart of tuberculosis incidence from 2015 to 2019

Figure 2 shows the distribution of pulmonary tuberculosis incidence rates in China (excluding Hong Kong, Taiwan, and Macau). It can be seen that Xinjiang, Tibet, and Qinghai witnessed the highest incidence rates, while Beijing, Tianjin, Shanghai, Jiangsu and other cities reported relatively fewest incidences. Generally speaking, the incidence rate is decreasing from west to east, and the incidence rate in inland areas is higher than that in coastal areas. The total incidence rate reported in China from 2015 to 2019 showed a decreasing trend year by year.

Global Moran Index Spatial Autocorrelation Analysis: It can be seen from Table 1, the tuberculosis incidence in various provinces in China from 2015 to 2019 has a significant spatial correlation. The Global Moran's I index ranges from 0.293 to 0.404, and all of the value P is 0.000, which means The Global Moran value is significant at 1% significance level. Therefore, the spatial correlation of tuberculosis incidence can be further studied.

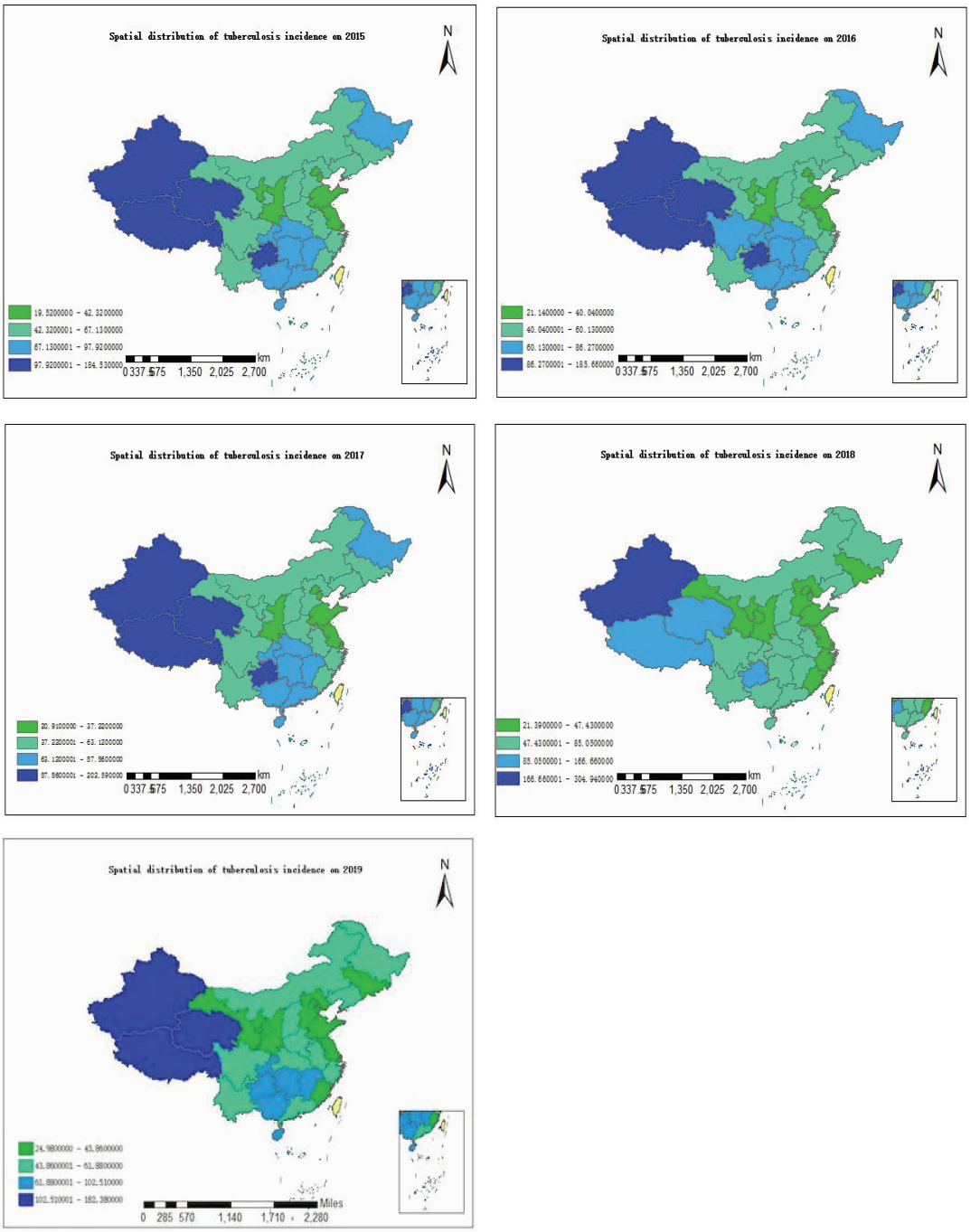
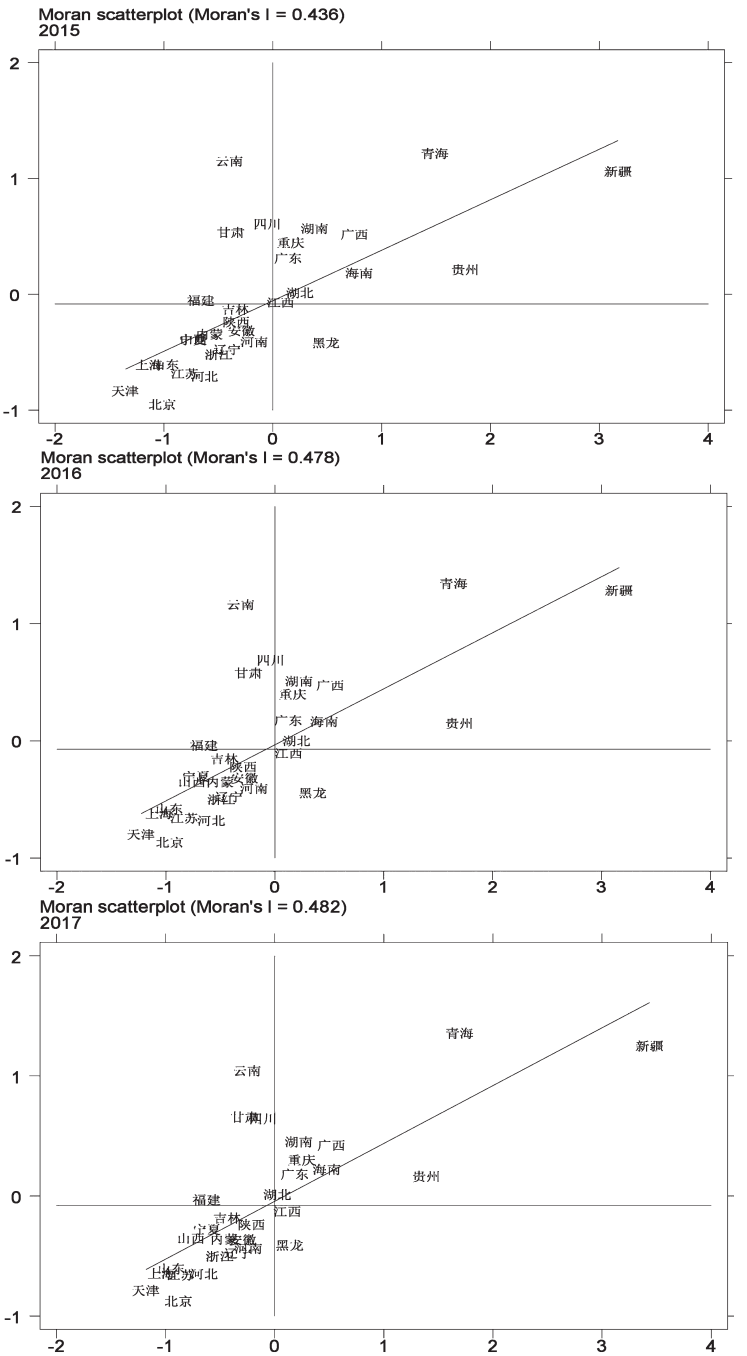


Figure 2 Spatial distribution of tuberculosis incidence from 2015 to 2019

Note: Based on the standard map production of GS (2016) No. 2923 of the standard map service website of the Ministry of Natural Resources, the details of the standard map are not modified.

Table 1 2015-2019 Moran Index of tuberculosis incidence

year	I	sd(I)	z	p–value*
2019	0.404	0.101	4.343	0.000
2018	0.293	0.085	3.833	0.000
2017	0.395	0.100	4.300	0.000
2016	0.387	0.102	4.128	0.000
2015	0.367	0.103	3.908	0.000



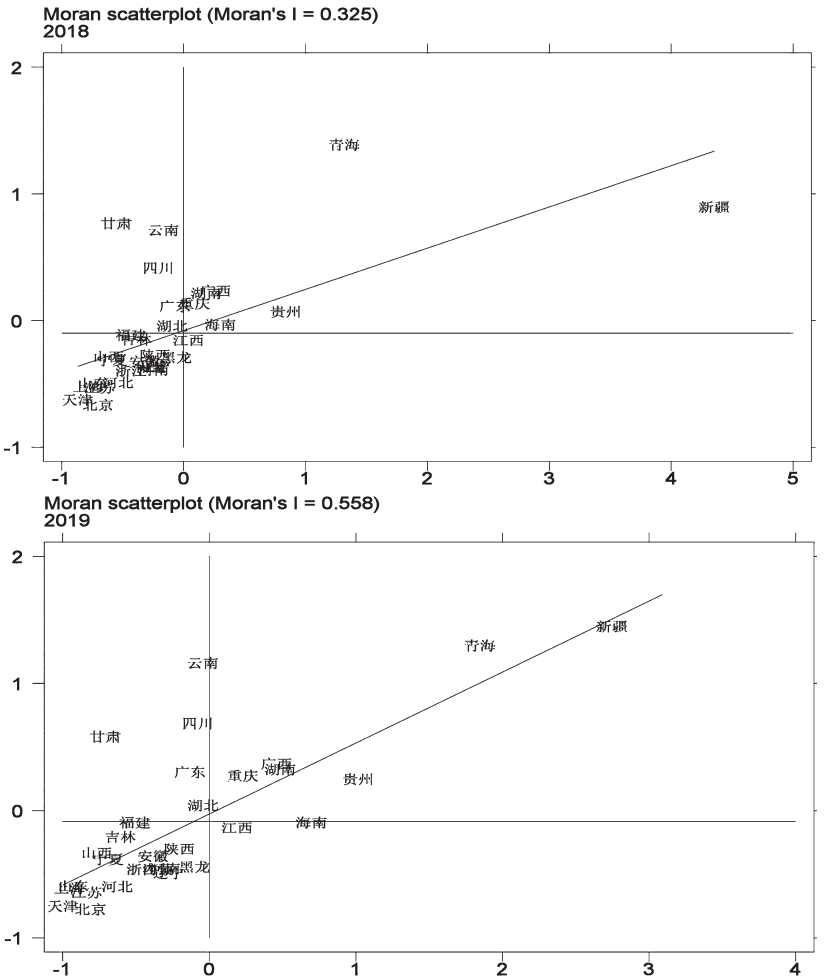


Figure 3 2015-2019 Moran scatter plot of tuberculosis incidence

Based on Moran scatter diagram, the distribution of provinces in the past five years can be shown in Figure 3 above. In the past five years, Xinjiang, Qinghai, Tibet, Guizhou, Guangxi, have long been in the first quadrant (HH), while Liaoning, Jilin, Shandong, Anhui, Zhejiang, Shanghai, Jiangsu, Hebei, Shanxi, Tianjin, Beijing, Henan, Ningxia, and Shaanxi have long been in the third quadrant (LL), and Gansu, Yunnan, and Sichuan have long been in the second quadrant (LH). During the five years from 2015 to 2019, the incidence rate in Xinjiang has been ranked first. The incidence of pulmonary tuberculosis shows obvious spatial clustering.

3.2 Results of cluster analysis of tuberculosis incidence

A cluster analysis of tuberculosis incidence from 2015 to 2019 was carried out, and the specific results are shown in Figure 4 below. The K-Means cluster (Figure 4 left) divides tuberculosis incidence in China into three groups. Xinjiang, Tibet, Guizhou and Qinghai were classified as the first group with the highest incidence rate; Heilongjiang, Henan, Hubei, Sichuan, Chongqing, Jiangxi, Guangdong, Guangxi and Hunan were classified into the second

group with median incidence rate; the remaining provinces were classified into the third group with the lowest incidence. Hierarchical clustering (Figure 4 right) divides the tuberculosis incidence in China into four categories, Xinjiang was classified into the first category; Tibet, Qinghai and Guizhou were classified into the second category; Heilongjiang, Chongqing, Hubei, Hunan, Jiangxi, Guangxi and Guangdong were classified into the third category, and the remaining provinces were classified into the fourth category. Comparing the two clustering methods, it can be concluded that the tuberculosis incidence in China can be roughly divided as: the highest incidence in the northwest, the medium incidence in the southwest, and the lowest incidence in the east. This is also in line with the conclusion that tuberculosis incidence in China is spatially clustered.

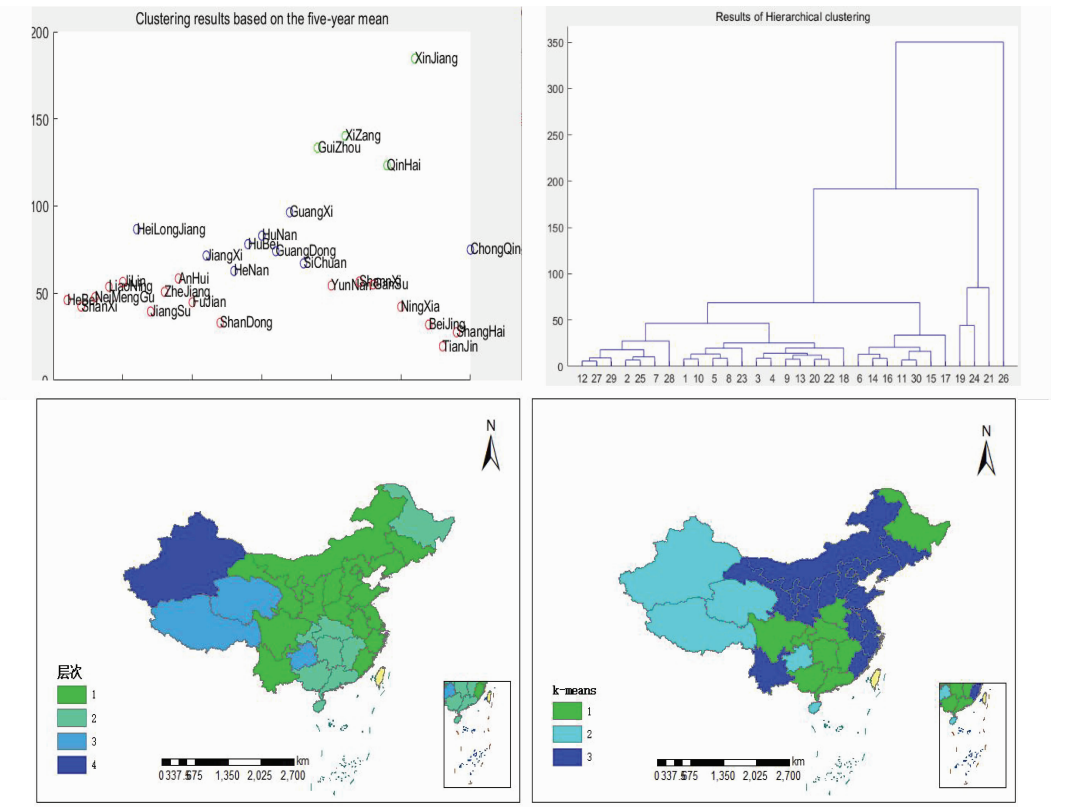


Figure 4 K-Means clustering and hierarchical clustering results

Note: Based on the standard map production of GS (2016) No. 2923 of the standard map service website of the Ministry of Natural Resources, the details of the standard map are not modified.

3.3 Analysis of influencing factors of tuberculosis incidence

In order to further analyze the factors affecting tuberculosis incidence, and study factors such as the level of economic development, climatic condition, and air quality in a comprehensive way, the least square method is adopted to analyze the influence of different factors on tuberculosis incidence. The paper chooses regional GDP per capita to measure the level of economic development(eco), annual average temperature to represent climatic conditions (tem), and annual average PM2.5 concentration to represent air quality (PM2.5). At the same time, the logarithm of all variables is taken to weaken the influence of heteroscedasticity to estimated results. The results are shown in Table 3 below.

Table 3 Regression result

Variable	OLS
PM2.5	–0.4132***(0.087)
eco	–0.6856***(0.083)
tem	–0.0903(0.084)
cons	13.37***(0.918)
Obs	155
R^2	0.4108

legend: * $p<.1$; ** $p<.05$; *** $p<.01$

It can be seen from the regression results that the level of economic development has a significant negative impact on tuberculosis incidence. This is because areas with a more developed economy can provide better medical services, which can play a better role in the prevention and treatment of local tuberculosis incidence. There are relatively low incidences in areas with higher average annual temperatures. This is perhaps due to the fact that areas with lower temperatures are mostly built with low-storey, crowded, and poorly-ventilated houses. Such kind of housing conditions may cause tuberculosis to spread faster. Therefore, paying attention to keeping warm, improving physical fitness and house ventilation can effectively prevent tuberculosis. Unexpectedly, the incidence rate in areas with higher PM2.5 concentrations is lower. Generally speaking, areas with more developed economies boost higher industrial levels, which may lead to poor local air quality. But economically developed areas can provide better medical services, which can reduce the impact of air quality on morbidity. In summary, the most influential factor in the tuberculosis incidence is the level of economic development.

China is a vast country with uneven economic development, climate conditions and air quality among different regions. In order to universalize the research results, according to the above cluster results, the sample is divided into the northwest region with the highest incidence, the southwest region with medium incidence, and the eastern region with low incidence to further study the influencing factors of tuberculosis incidence. The specific results are shown in Table 4 below.

Table 4 Regression results by region

Variable	East	Northwest	Southwest
PM2.5	–0.2003***	0.2072	0.0959
	(0.075)	(0.202)	(0.096)
eco	–0.7144***	–0.3947	–0.0811
	(0.077)	(0.534)	(0.114)
tem	0.2000*	–0.2469	0.1788***
	(0.085)	(0.181)	(0.058)
Cons	9.2791***	8.6808	4.2133***
	(0.781)	(4.749)	(1.146)
Obs	100	20	40
R2	0.3999	0.1828	0.2099

4 Discussion

According to the first law of geography: "the closer, the more similar", things close in space share a certain similarity. Spatial statistical analysis, a quantitative study on geospatial phenomena, is a kind of statistical analysis method based on spatial data. Spatial analysis mainly makes the joint analysis of spatial data and spatial models to mine the potential information behind spatial targets. Spatial calculation and analysis of many specific tasks can be performed by combining the spatial data and attribute data of spatial targets. The spatial data in spatial statistical analysis is not isolated, that is, the data has a certain correlation in space. In epidemiological research, spatial statistics and geospatial information are chosen to explore the role of geospatial information in the spread of infectious diseases, which can provide reasonable suggestions and countermeasures for the prevention and control of infectious diseases. Therefore, it is necessary to report areas and times with a higher tuberculosis incidence to decision-makers since it can help them to take corresponding measures and better prevent the spread of tuberculosis.

From the above analysis results, there may be a seasonal trend in the transmission of tuberculosis. Based on the data from 2015 to 2019, the number of reported cases peaked in March each year and then gradually declined. The number of reported cases is the least in February each year and rose sharply from February to March. There was another small peak in November, but much smaller than that in March, which may have a certain relationship with the transmission mode and characteristics of tuberculosis. According to the incidence rates reported by provinces, the provinces with higher incidence rates in the past five years are Xinjiang, Tibet, Qinghai, Guizhou, and Guangxi. Provinces with relatively low incidence rates include Tianjin, Beijing, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, etc. Overall, the incidence rate in inland areas is higher than that in coastal areas, and the incidence rate in western areas is higher than that in eastern areas. This may be related to climatic conditions, geographical environment, and living customs. In addition, the incidence rate reported in China is declining year by year, indicating that the prevention and control of tuberculosis in China are achieving results and getting better year by year.

In the autocorrelation analysis of tuberculosis incidence from 2015 to 2019, the Global Moran's I index is all positive, and both are significant at the 1% significance level, indicating that the spatial distribution of tuberculosis incidence has a positive correlation, among which the spatial correlation in 2019 is the strongest. On the one hand, Xinjiang, Qinghai, and Tibet have long been in the first quadrant with high agglomeration provinces; provinces with low agglomeration in the third quadrant for a long time include Tianjin, Beijing, Shanghai, Jiangsu, Shandong, Zhejiang and other eastern coastal cities. In the evolution of high-incidence areas of tuberculosis from 2015 to 2019 in the past five years, high-accumulation areas are dominated by Xinjiang, Tibet, and Qinghai, while low-accumulation areas are dominated by Tianjin, Beijing, Jiangsu, and Shandong. It can be concluded that the natural environment, such as climate, sunlight, and rainfall, play a major role in the spread of tuberculosis.

In the cluster analysis of tuberculosis incidence from 2015 to 2019, tuberculosis incidence in China can be divided into four categories. The first category: Xinjiang, the highest incidence; the second category: Tibet, Qinghai and Guizhou, the slightly higher incidence rate; the third category: Heilongjiang, Jiangxi, Hubei, Hunan, Sichuan, Chongqing, Guangdong and Guangxi, the moderate incidence; and the fourth category: other provinces and cities, the slightly lower incidence. In the analysis of the influencing factors of tuberculosis incidence,

the higher level of economic development, the lower tuberculosis incidence, and the level of economic development in the eastern region have a greater impact on the tuberculosis incidence. Incidence in the northwest and east is affected by temperature. The tuberculosis incidence in the western region is mainly affected by air quality, so the western region can consider reducing the tuberculosis incidence by improving air quality.

5 Conclusions

We use stata software to draw local Moran scatter plots for 31 provinces in China (excluding Hong Kong, Macau, and Taiwan). There also are certain limitations in this paper. If there is no land connection between a certain area and other areas, the weight of the area is unavailable when calculating the spatial weight matrix. Although there is no land connection between Hainan Province and other provinces, tuberculosis incidence reported by Hainan Province has been at a high level throughout the country, so it is also taken into the calculation range. In addition, we not only analyze the spatial aggregation of high-frequency morbidity, but also consider other factors that contribute to high-frequency transmissions, such as bad climate, poor natural environment, and an undeveloped economy. The level of economic development and air quality has significant influence and the most influential one is economic development. The high level of economic development and medical conditions in the region has significantly reduced the tuberculosis incidence in the east, and air quality is also one of the important factors affecting the incidence in the northwest and southwest regions. Finally, there are a few influencing factors considered in this paper, and if we want to further explore the reasons affecting the tuberculosis incidence in China, we can consider more influencing factors and do more comprehensive research.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

Acknowledgments

There is no project funding support with this paper.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geogr Anal, 27* (2), 93–115.

César, P., Graça, C., Fortes, P., & Guerreiro, V. (2021). Pulmonary tuberculosis – Still a reality. *Medicina Clínica Práctica, 4* (1), 100164.

Dimala, C. A., Kadia, B. M., & Hansell, A. (2020). The association between ambient air pollution and pulmonary tuberculosis: A systematic review protocol. *Environmental Evidence, 9* (1), 29.

Ejemot–Nwadiaro, R. I., & Nja, G. (2020). Socio–demographic and nutritional status correlates in pulmonary tuberculosis patients in Calabar, Nigeria. *Asian Journal of Medicine and Health, 18* (10), 85–98.

Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. *Geogr Anal, 24* (3), 189–206.

Gregorius, H. (2004). The isolation approach to hierarchical clustering. *Classification, 21* (1), 51–69.

Hussain, Z. (2020). Investigation of a cluster of pediatric pulmonary tuberculosis cases in Gilgit–Baltistan (GB) Pakistan 2019. *International Journal of Infectious Diseases, 101* (S1), 223–224.

Kumar, R., Krishnan, A., Singh, M., Singh, U. B., Singh, A., & Guleria, R.. (2020). Acceptability and adherence to peanut–based energy–dense nutritional supplement among adult malnourished pulmonary tuberculosis

patients in Ballabgarh block of Haryana, India. *Food and Nutrition Bulletin, 41* (4), 438–445.

Liu, M. Y., Li, Q. H., Zhang, Y. J., Ma, Y., Liu, Y., Feng, W., et al. (2018). Spatial and temporal clustering analysis of tuberculosis in the mainland of China at the prefecture level, 2005– 2015. *Infect Dis Poverty, 7* (1), 106.

Mai, X., Cheng, J., & Wang, S. (2019). Research on semi supervised K–means clustering algorithm in data mining. *Cluster Computing, 22* (2), 3513–3520.

Ord, J. K., & Getis, A. (2010). Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr Anal, 27* (4), 286–306.

Peng, A. Z., Yang, A., Li, S. J., Qiu, Q., & Chen, Y. (2020). Incidence, laboratory diagnosis and predictors of tracheobronchial tuberculosis in patients with pulmonary tuberculosis in Chongqing, China. *Experimental and Therapeutic Medicine, 20* (6), 174.

Šulc, Z., & Řezanková, H. (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification, 36* (1), 58–72.

Talpur, S. K., Kumar, M., Abbass, A., Jan, N. A., Lohano, K. C., & Menmon, I. A. (2020). Comparison of anti–tuberculosis treatment outcomes of pulmonary tuberculosis in current, ex and non smokers. *Journal of Pharmaceutical Research International, 32* (32) 32–38.

Zhao, Y. P., & Zhou, X. L. (2021). K–means clustering algorithm and its improvement research. *Journal of Physics: Conference Series, 1873* (1).