Original Research Article

The Research on Causes of Megacity Problem Based on Spatial Disequilibrium

QI Zi-xiang[1,] ,LV Yong-qiang[2],WANG Ya-xin*

1 School of Economics，Beijing Wuzi University，Beijing 101149，China；

* **Corresponding author:** *QI Zi-xiang*

Abstract: By using spatial models with lag and error and through many experiments, this paper uses distance threshold which can maximize spatial autocorrelation, instead of setting spatial weight matrix by adjacent boundaries, making measurement better suit the features of our country's cities distribution to explore the reasons for big city disease empirically. It finds that distribution of public service distribution and spatial non-equilibrium are factors for excessive population agglomeration and therefore big city disease, and provides solutions for the factors mentioned above.

Keywords: Spatial disequilibrium; Spatial weights; Models with lag and error; Generalized spatial two stage least squares estimation.

1. Introduction

The megacity problem manifests itself in the excessive concentration of population in core cities, which exceeds the upper limit of carrying capacity, thus triggering a series of urban governance problems such as traffic congestion, environmental pollution, high housing prices and the heat island effect. General Secretary Xi Jinping has clearly pointed out that urban construction is a matter of urban governance, and that it is necessary to rationalize the layout of production, living and ecology, and improve the livability and sustainability of urban development. Obviously, the management of megacity problem by local governments has become a top priority of urban governance and an urgent need to improve the well-being of the people. The paper aims to use spatial econometrics to empirically investigate the causes of megacity problem, so as to provide a decision-making basis for China's regional development strategy and urban planning, and to serve the National 14th Five-Year Plan.

2. Literature Review

The use of spatial econometrics to study spatial problems can be traced back to the work of Paelinek's 1967 report at the annual French Regional Science Conference[1]. After Moran (1947) proposed the 0-1 linking matrix to represent spatial correlation, he proposed Moran's I statistic in 1950 to measure spatial autocorrelation[2]. Geary (1954) gave another measure of spatial dependence, Geary's C[3]. Cliff and Ord (1972) suggested that Moran's I could be used to test for the presence of spatial autocorrelation effects in the residuals from least squares regression[4]. Ord (1975) proposed a spatial error model and a spatial lag model and gave a maximum likelihood estimation method[5]. Paelinck and Klaassen in 1979 proposed five important research areas for spatial econometric modeling: spatial dependence, spatial asymmetry of relationships, explanatory factors of other spatial units, ex-ante and ex-post factors, and spatial dependence, differences in ex ante and ex post interactions, and spatially explicit modeling[6]. Anselin (1988, 2006) argued that the development of spatial econometrics should be different from spatial statistics, focusing more on the measurement theory of spatial modeling, estimation, and testing, and derived and summarized the maximum likelihood estimation of generalized spatial models[7,8]. Pace (1997) generalized a fast computational method for maximum likelihood

Received: 12 September 2023 Accepted: 23 September 2023 Available online: 19 October 2023

estimation based on LU decomposition[9]. Barry (1999) gave a Monte Carlo method based on the logarithmic determinant of sparse matrices[10]. Pace (2004) discussed the Chebyshev approximation of the logarithmic determinant of sparse matrices[11]. Zhang (2007) discussed the logarithmic determinant approximation of Gaussian processes[12]. Bivand (2013) demonstrated the Jacobi determinant based on Gaussian spatial autoregressive modeling computational method[13]. Anselin (1980) used spatial two-stage least squares to estimate the spatial lag model[14]. The generalized method of moments estimation was first proposed by Kelejian (1998, 1999) [15,16], and further developed by Kelejian (2010), and refined by Drukker (2013) [17,18]. Hepple (1979) discusses Bayesian estimation of spatial econometric models[19]. Lesage (1997) applies Markov chain Monte Carlo methods and Gibbs sampling to Bayesian estimation of spatial econometric models[20]. Thanks to the computational simplification, Bayesian estimation has been widely used. For example, Wang (2012) discusses land use change based on Bayesian estimation of a dynamic spatial discrete choice model[21]. Tests of spatial econometric models are not only Moran's (1950) proposed Moran's I statistic proposed by Moran (1950), Geary's C statistic proposed by Geary (1954), the Getis (1995) proposed G statistic[22], the Wald test, the Lagrange multiplier test, and the likelihood ratio test are also applicable. These classical tests have also been improved with the development of spatial econometrics. Anselin (1997) constructed a robust Lagrange multiplier test statistic that greatly facilitates model setting in practical applications[23]. Baltagi (2001) relies on Lagrange multipliers to test whether the function format has been set incorrectly[24]. Anselin (2001) discussed the application of the Lagrange multiplier test to the case of autocorrelation of other spatial errors[25]. Lauridsen (2006) proposed a unit root test based on the LM test[26]. Kelejian (2008) discussed the testing of non-nested hypotheses[27]. In addition, the applicability of other tests on various models is also discussed. For example, Amaral (2014) discusses the nature of Moran's I in the spatial Tobit model[28]. The key to whether the above various spatial econometric methods can correctly estimate and test the spatial effects lies in choosing appropriate spatial weight matrices for different research problems and parameterizing the spatial autocorrelation process according to the appropriate matrix setting method, so as to uncover the spatial disequilibrium phenomena hidden in the economic problems.

3. Hypothesis on the Causes of Megacity Problem

Taking Beijing as an example of a megacity, this paper investigates whether there are spatial imbalances of the above three factors in Beijing through spatial statistics, starting from the three dimensions of market (circulation economy), society (public services) and employment (livelihood issues), and verifies them with a spatial econometric model. And combined with spatial econometric modeling, it is verified.

3.1 Hypothesis 1: Spatial Imbalance in the Layout of Circulation Economy

People's production and life can not be separated from circulation, circulation economy reflects the trade status and prosperity of a city. Circulation business is divided into wholesale and retail. Due to the space limitation, the thesis uses the density of wholesale enterprises to reflect the market trade condition, using the POI vector data of 1,835 wholesale enterprises in Beijing in 2012, and utilizing the kernel density estimation method to investigate the spatial pattern of circulation economy in Beijing, and the estimation results are shown in Figure 1. The results now show that the wholesale enterprises in Beijing present a spatial pattern of clustering along the railroad transportation lines. The spatial imbalance in the distribution of the circulation economy leads to the spatial difference in the size of the circulation economy between districts.

This paper develops a first hypothesis about the megacity problem. Hypothesis 1: There is a spatial imbalance in the distribution of the circulation economy in Beijing. This imbalance may be one of the causes of the megacity problem.

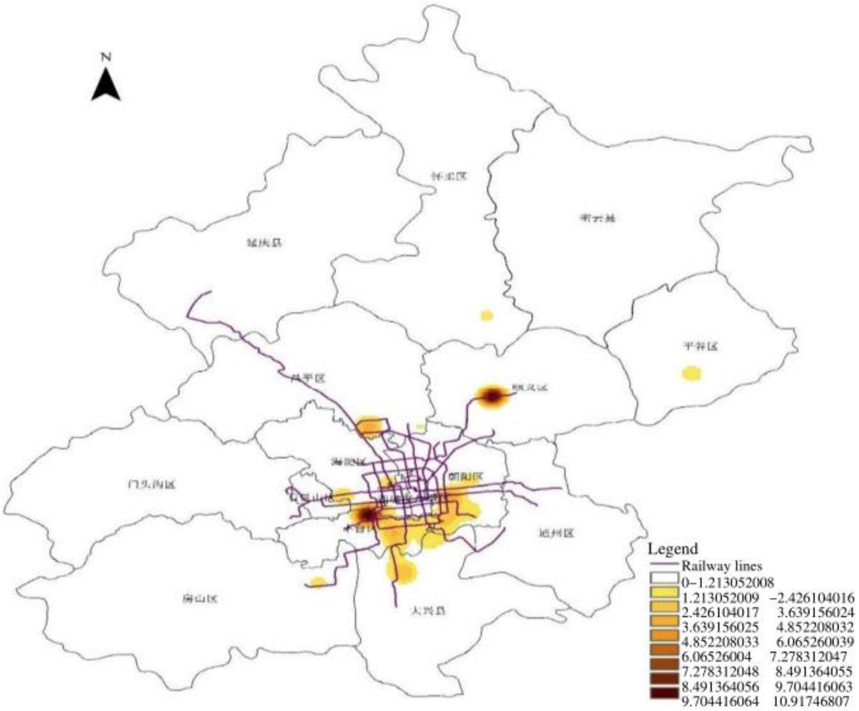

Figure 1 Spatial Concentration Pattern of Wholesale Enterprises in Beijing (2012).

3.2 Hypothesis 2: Spatial Imbalance in the Layout of Public Services

This paper deals with public services in the field of social development, mainly including education, health care, culture, sports, public security, social welfare and social assistance. The paper draws on the local G_i statistic (Eq.1) proposed by Getis (1995) [22], and uses data on the daily outpatient volume of 67 tertiary hospitals in Beijing in 2010 to conduct a spatial hotspot analysis of the public service organizations represented by hospitals in Beijing, and to explore the spatial pattern of the public service sector.

$$G_i = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x} \sum_{j=1}^{n} w_{ij}}{\delta \sqrt{\left[\frac{\left[n \sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij} \right)^2 \right]}{n-1} \right]}} \tag{Eq.1}$$

where x_j is the daily outpatient volume of the j th hospital, w_{ij} is the spatial weight between hospital i and hospital j. The paper constructs a spatial weighting matrix with a distance threshold of $d = 2000$ meters, and n is the number of hospitals. $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$, $\delta = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{x}\right)^2}$. The g_i-statistic z-score is often used as a tool for "hot spot" analysis.The formula for the g_i-statistic z-score is $zg_i = \frac{g_i - E\left(g_i\right)}{\sqrt{V\left(g_i\right)}}$. A high and positive g_i-statistic z-score indicates the presence of high-value clusters or hot spots. Conversely, a low and negative g_i-statistic z-score indicates the presence of low-value agglomerations or cold spots. The paper tests the significance of the g_i-statistic through the z-score, which follows a standard normal distribution when the sample size is sufficient.

The p-value denotes the probability that the g_i-statistic z-score will be returned returned with a confidence level encompassing 90%, 95%, or 99%. For hotspot analysis, the p-value indicates the probability that the observed spatial pattern was created by some random process. When p-value is very small, it means that the observed spatial pattern is unlikely to have arisen from a stochastic process (a small probability event), and therefore the paper can reject the original hypothesis that the g_i-statistic is not significant. The results of the hot spot analysis are shown in Table 1 and Figure 2.

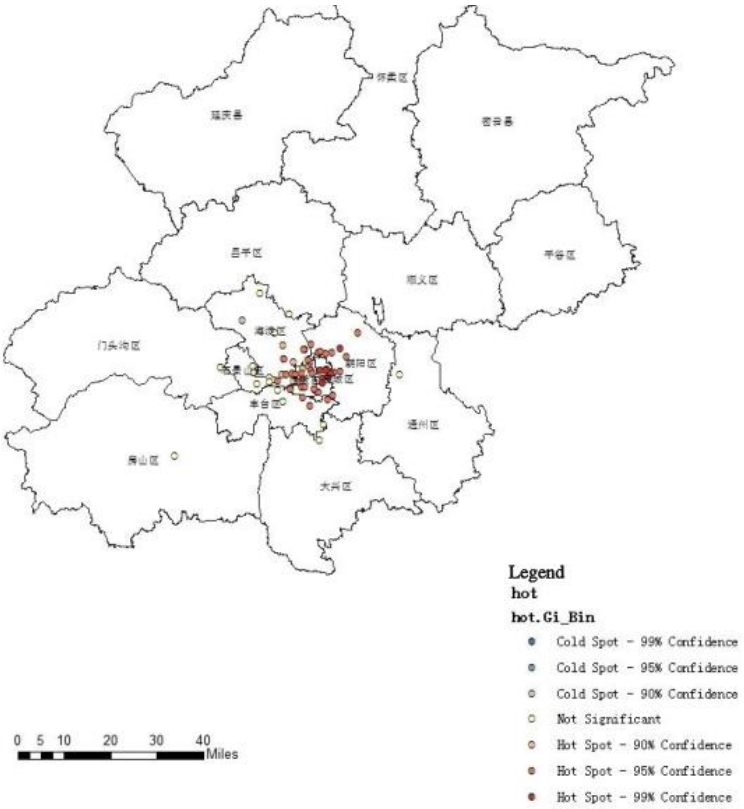
延庆县
怀柔区
密云县
昌平区
顺义区
平谷区
门头沟区
海淀区
石景山区
朝阳区
丰台区
通州区
房山区
大兴区
Legend
hot
hot.Gi_Bin
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
0 5 10 20 30 40
Miles

Figure 2 Hotspot analysis of tertiary hospitals in Beijing.

Among the 67 tertiary hospitals in Beijing, 50 hospitals are the hotspots for people to see the doctor, as can be clearly seen in Figure 2, which shows that they are clustered in Beijing's Chaoyang District, Xicheng District, Dongcheng District and Haidian District. Public medical resources are clustered in the core urban areas of Beijing, and the spatial imbalance is very obvious. In addition to the above four districts, the other 12 districts and counties in Beijing do not have a single hotspot hospital for people to see a doctor. The spatial imbalance in public services leads to inter-district disparities in public service capacity.

This paper develops a second hypothesis about the megacity problem. Hypothesis 2: There is a spatial imbalance in public services in Beijing. This imbalance may be one of the causes of the megacity problem.

3.3 Hypothesis III: Spatial Disequilibrium in Employment Opportunities

In order to depict the spatial change trend of the employed population in Beijing, this paper applies the LOESS method of non-parametric estimation to the share of the employment density of each street and township in the total employment density of Beijing, and fits the law of the change in the distance from the city center of the employment density of each street and township through the curve. The smoothing coefficient of the LOESS curve is set to 0.5 for comparison purposes. By comparing the fitted LOESS curves in 2008 and 2013, as shown in Figure 3 (the horizontal coordinate indicates the distance to the city center, and the vertical coordinate indicates the share of employment density), we can find the trend characteristics of population agglomeration and diffusion in different spatial scopes: during the period of 2008-2013, the employment density in Beijing further increased in the city, and the trend of agglomeration to the city center remained unchanged, which, from a side This reflects the unbalanced spatial distribution of employment opportunities.

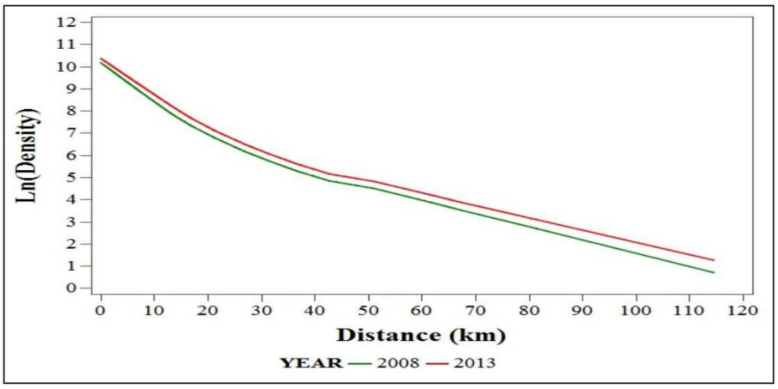

Figure 3 Employment density LOESS curve in Beijing, 2008 and 2013.

Table 1 Hot spot analysis of daily outpatient volume in tertiary hospitals in Beijing (2010).

Hospital Name	Daily Outpatient Visits	Bureau g_i Statistic z Score	P-value	Distribution Type
Beijing Union Medical College Hospital	12000	2.6911***	0.0071	hot spot
General Hospital of the Chinese PLA	9315	2.0352**	0.0418	hot spot
Beijing University Third Hospital	9589	2.0689	0.0385	hot spot
Beijing Tongren Hospital	4250	2.3856	0.0170	hot spot
Beijing Children's Hospital	6301	2.1175	0.0342	hot spot
Fu Wai Hospital	342	2.1175	0.0342	hot spot
China-Japan Friendship Hospital	6100	2.2680	0.0233	hot spot
Beijing University People's Hospital	6669	1.8911	0.0586	hot spot
Guang'anmen Hospital	7200	2.1100	0.0348	hot spot
Beijing Chinese Medicine Hospital	6000	2.6911	0.0071	hot spot
Beijing Tiantan Hospital	1500	2.1725	0.0298	hot spot
Beijing University First Hospital	7000	2.3810	0.0172	hot spot
Beijing Anzhen Hospital	3964	2.1957	0.0281	hot spot
Beijing Jishuitan Hospital	3000	1.8911	0.0586	hot spot
Xiyuan Hospital	4000	1.7404	0.0817	hot spot
Beijing Maternity Hospital	2795	1.9892	0.0466	hot spot
Beijing Chaoyang Hospital	6849	2.5056	0.0122	hot spot
Beijing University Stomatological Hospital	3205	1.8911	0.0586	hot spot
Dongzhimen Hospital	3992	2.8379	0.0045	hot spot
Xuanwu Hospital	5000	2.1100	0.0348	hot spot
Tumor Hospital	1644	2.3525	0.0186	hot spot
The First Hospital of Beijing University	7000	2.3810	0.0172	hot spot
The First Affiliated Hospital of PLA	135	2.1641	0.0304	hot spot
Beijing Hospital	132	2.6296	0.0085	hot spot
Beijing University Cancer Hospital	767	1.9187	0.0550	hot spot
General Hospital of the Air Force	3562	1.9786	0.0478	hot spot
Beijing Jishuitan Hospital	3000	2.0482	0.0405	hot spot
Beijing Friendship Hospital	8000	2.3856	0.0170	hot spot
Xiyuan Hospital	4000	2.8379	0.0045	hot spot
The 302 Hospital of the Chinese PLA	58	1.1503	0.2499	insignificant
Children's Hospital	5886	2.8836	0.0039	hot spot
Armed Police General Hospital	4500	1.3291	0.1838	insignificant
The 306 Hospital of the Chinese PLA	1800	2.2317	0.0256	hot spot
Beijing Stomatological Hospital	2000	2.1725	0.0298	hot spot
Dongfang Hospital	91	2.3525	0.0186	hot spot
Naval General Hospital of the PLA	2100	2.1175	0.0342	hot spot
Beijing Anding Hospital	425	1.9673	0.0491	hot spot
Second Artillery General Hospital	1095	1.9673	0.0491	hot spot
The 309 Hospital of the Chinese PLA	104	-0.5880	0.5564	insignificant
Beijing You'an Hospital	682	2.1716	0.0298	hot spot
Beijing Ditan Hospital	2600	2.0647	0.0389	hot spot
Fuxing Hospital	118	2.1175	0.0342	hot spot
Plastic and Reconstructive Surgery Hospital	33	-0.6429	0.5202	insignificant
The Sixth Hospital of Beijing University	30	1.9673	0.0491	hot spot
Beijing Tongren Hospital Yizhuang Campus	49	-0.2403	0.8100	insignificant
Wangjing Hospital	3500	2.5795	0.0098	hot spot
Xiyuan Hospital	4000	1.0914	0.2750	insignificant

Beijing Chaoyang Hospital West Campus	1300	-1.0557	0.2910	insignificant
Aerospace Center Hospital	1682	1.0017	0.3164	insignificant
Beijing Huilongguan Hospital	45	-0.3928	0.6944	insignificant
Beijing Huaxin Hospital	2329	2.4562	0.0140	insignificant
Beijing Chaoyang Hospital West Campus	1300	2.0437	0.0409	hot spot
Civil Aviation General Hospital	2400	2.6256	0.0086	hot spot
Beijing Boai Hospital	489	1.9820	0.0474	hot spot
Shougang Hospital of Beijing University	2742	-0.1004	0.9199	insignificant
Beijing Chest Hospital	463	-0.7620	0.4460	insignificant
Beijing Sanbo Brain Hospital	9	2.0482	0.0405	hot spot
Ophthalmology Hospital	800	0.7259	0.4678	hot spot
The 306 Hospital of the Chinese PLA	1800	2.4430	0.0145	hot spot
The 305 Hospital of the Chinese PLA	1000	2.2929	0.0218	hot spot
Beijing Aerospace General Hospital	1700	1.4148	0.1571	insignificant
The 305 Hospital of the Chinese PLA	1000	-0.6796	0.4967	insignificant
Beijing Xuanwu District Hospital	1200	2.2861	0.0222	hot spot
Beijing Electric Power Hospital	1600	2.1697	0.0300	hot spot
Beijing Geriatric Hospital	978	-1.9193	0.0549	hot spot
Beijing Coal Group General Hospital	548	-1.4061	0.1596	insignificant
Beijing Aerospace General Hospital	1700	-1.2595	0.2078	insignificant

The paper concludes with a third hypothesis on megacity problem. Hypothesis 3: There is a spatial imbalance in employment opportunities in Beijing. This imbalance may be one of the causes of the megacity problem.

4. Empirical Tests on the Causes of Megacity Problem

Through the spatial analysis of Beijing, this paper draws three hypotheses on the causes of megacity problem. Based on the above hypotheses, we empirically test whether the hypotheses are valid based on the national perspective.

4.1 Data Sources and Variable Interpretation

Currently, most of the spatial measurement studies in China use faceted domain data to set the spatial weight matrix with boundary adjacency. The disadvantage of this research paradigm is that provincial, municipal, and county data cover both cities and villages, and cannot clearly portray the attributes and characteristics of cities in space.

The data in this paper comes from China Urban Statistical Yearbook 2015, 289 cities at prefecture level and above in mainland China were selected as statistical samples, vector point data were utilized, the spatial weight matrix setting method of boundary adjacency was abandoned. Setting the spatial weight matrix based on the distance threshold that leads to the most significant spatial autocorrelation effect of the explanatory variables. Taking the population density of the municipal district at the end of the year of the city as the explanatory variable (popden), portraying the degree of population agglomeration of the city. Taking the number of industrial enterprises above large scale in the municipal jurisdiction as the explanatory variable (enter), reflecting the employment opportunities. The total retail sales of consumer goods in the city as the explanatory variable (sale), which represents the scale of the distribution economy. Using the number of hospitals and health center beds in the municipal area as the explanatory variable(hospi), representing the capacity of public services. The causes of megacity problem are empirically examined. The explanations of the variables are given in Table 2 and the variables are described in Figure 4.

Table 2 Explanation of variables.

	Variable Name	Meaning	Alphabetical representation
Explained Variables	Population density of the city at the end of the year	Degree of population concentration	Inpopden
Explanatory Variables	Number of Industrial Enterprises Above Scale	Employment opportunities	lnenter
Explanatory Variables	Total retail sales of consumer goods	Size of the circulation economy	lnsale
Explanatory Variables	Number of beds in hospitals and health centers	Public service capacity	lnhospi

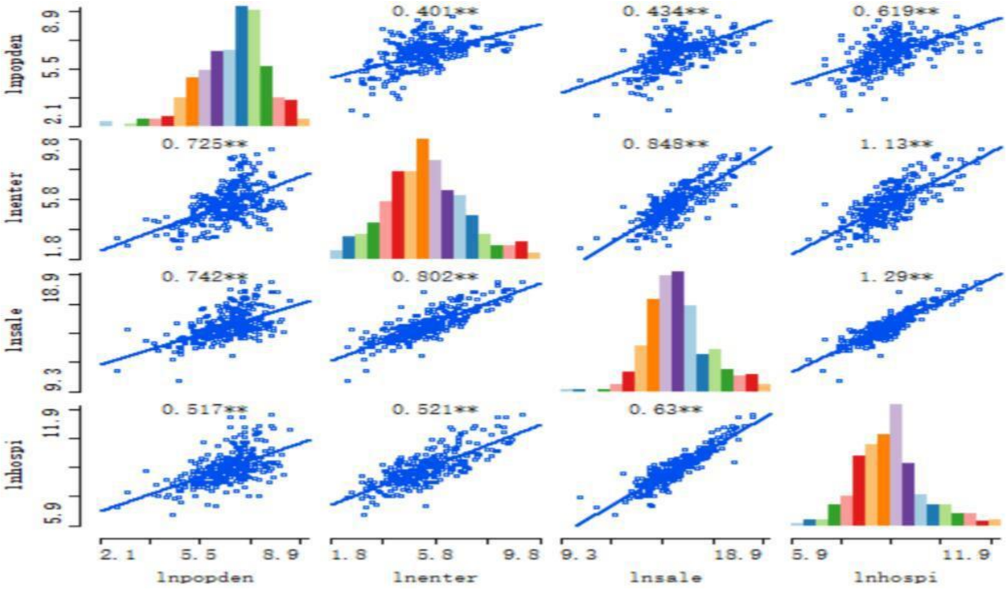

Figure 4 Variable description with scatterplot.

4.2 Spatial Autocorrelation Test

Does the regression process use classical measures, or spatial measures? A spatial autocorrelation test is required.

(1) Full domain spatial autocorrelation test defined by distance.

Moran (1950) [8] proposed the full domain autocorrelation index Moran's I to measure the correlation between variables according to the spatial assignment status. The value of the Moran's Index I of the global autocorrelation is between -1 and 1. A value greater than 0 indicates a positive correlation, which means that similar values are spatially close together, and the closer it is to 1, the stronger the correlation is, which can also be interpreted as clustering. A value less than 0 indicates negative correlation, which means that similar values are close in space, the closer to -1, the stronger the spatial differentiation. If the value is close to 0, it means that the variables are randomly distributed and there is no spatial autocorrelation. The formula of Moran's I is shown in equation (2).

$$I=\frac{N\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n}w_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)}{\delta^2\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}} \qquad \text{(Eq.2)}$$

x_i is the population density of the municipal district at the end of the year of the ith city (the explanatory variable), and n is equal to the total number of cities. w_{ij} is the spatial weight matrix. $w_{ij}=\begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$. The paper measures a range of distances for the domain-wide spatial autocorrelation index of urban population density, and selectively creates line plots of these distances and their corresponding z-scores. The z-scores reflect the degree of spatial clustering, with statistically significant peak z-scores corresponding to the spatial autocorrelation that promotes the most pronounced clustering of spatial processes as well as the distance threshold d, as in Figure 5.

The full domain spatial autocorrelation indices and z-scores obtained by constructing the spatial weight matrix according to the distance d in the ten experiments are listed in Table 3. The statistically significant peak z-score appeared in the third experiment, corresponding to the distance threshold $d=1198300.76$ meters, the z-statistic score $d=8.015472$, the highest in the ten experiments, and greater than the standard normal distribution at the significance level of 0.05 critical value of 1.96, which is the highest in the ten experiments.

From this, it is determined that there is spatial autocorrelation, i.e. spatial dependence, in the distribution of urban population density in China, and the degree of spatial autocorrelation is the most significant at the distance threshold $d = 1198300.76$ meters, with the spatial autocorrelation index Moran's $I = 0.047512$. After that, the z-score decreases rapidly with increasing distance and the degree of spatial autocorrelation (Moran's I-value) also decays gradually with increasing distance. This also confirms the first law of geography. The paper normalized the results of the third experiment to fall in a Cartesian coordinate system, as shown in Figure 6.

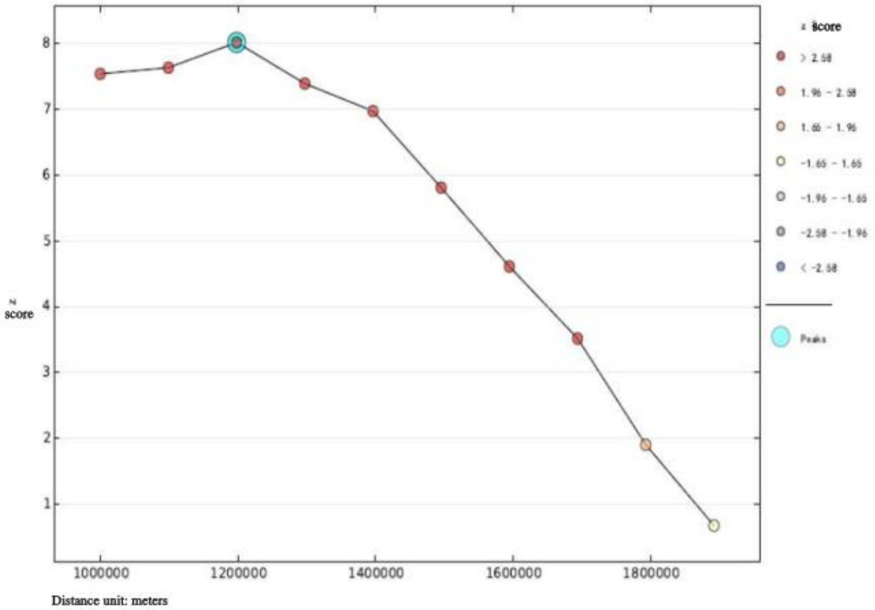

Figure 5 Full spatial autocorrelation defined by distance.

Table 3 Summary of Moran's 1 for full spatial autocorrelation by distance.

Number of experiments	Distance d (meters)	Moran's I index value	Variance	z Score	P-value
1	1000000.00	0.060298	0.000072	7.538625	0.000000
2	1099150.38	0.056869	0.000063	7.632514	0.000000
3	1198300.76	0.047512	0.000040	8.015472	0.000000
4	1297451.15	0.039355	0.000034	7.390805	0.000000
5	1396601.53	0.030939	0.000024	6.972150	0.000000
6	1495751.91	0.021353	0.000018	5.809503	0.000000
7	1594902.29	0.014506	0.000015	4.611839	0.000004
8	1694052.67	0.008521	0.000012	3.518401	0.000434
9	1793203.05	0.002151	0.000009	1.904204	0.056884
10	1892353.44	-0.001672	0.000007	0.676494	0.498727

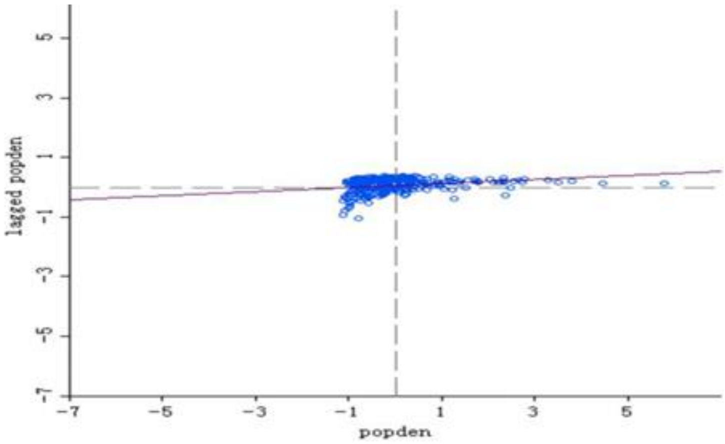

Figure 6 Spatial autocorrelation of population density in China's cities across the region.

In Figure 6, the slopes of the curves in the Cartesian coordinate system are both the spatial effect coefficients ρ (ρ =Moran's I) in the Spatial Autoregressive Model (SAR) and the General Spatial Autocorrelation Model (SAC), and the index of the global spatial autocorrelation of urban population density. The quadrants, on the other hand, represent the point distribution of cities in the localized spatial autocorrelation.

(2) Local spatial autocorrelation test defined by distance.

The global spatial autocorrelation index can only indicate that variables with similar values exhibit spatial clustering, it cannot indicate whether this clustering consists of high values, or low values. Anselin (1995) [29] proposed the Local Moran's I. (Local Moran index), or LISA (Local Indicator of Spatial Association), is a local indicator that describes spatial associations and is used to test whether similar or dissimilar observations cluster together in a local area. The local Moran index I_i of city i, which measures the degree of association between city i and his neighboring cities, is defined as equation (3).

$$Moran'sI_i = \frac{\left(x_i - \bar{x}\right)}{\delta^2} \sum_j w_{ij} \left(x_j - \bar{x}\right) \qquad \text{(Eq.3)}$$

$$\delta^2 = \frac{1}{n} \sum_i \left(x_i - \bar{x}\right)^2, \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$w_{ij} = \begin{cases} 1, & When\ the\ distance\ d_{ij}\ between\ city\ i\, and\ city\ j < threshold\ d \\ 0, & When\ the\ distance\ d_{ij}\ between\ city\ i\ and\ city\ j > threshold\ d \end{cases}$$

The paper sets the spatial weight matrix in the local autocorrelation index Moran's I in the following way. If the distance between two cities (Euclidean distance) is less than or equal to $d = 1198300.76$ meters, record $w_{ij} = 1$, otherwise record $w_{ij} = 0$.

The local spatial autocorrelation of each quadrant in Figure 6 is plotted on the map. As shown in Figure 7. Cluster: High represents the city point data in the first quadrant, indicating that cities with the same high population density are clustered together. Cluster: Low represents the city point data in the third quadrant, indicating that cities with the same low population density are clustered together. High Outlier represents city point data in Quadrant four, indicating that cities with high population density are adjacent to cities with low population density. Low Outlier represents city point data in Quadrant two, indicating that cities with low population density are adjacent to cities with high population density. White dots indicate cities that did not pass the significance test. The local autocorrelation test shows that China's urban population is clustered on the east side of “Hu Huan-yong line”. The population distribution of Beijing-Tianjin-Hebei, the Yangtze River Delta and the Central Plains urban agglomerations shows high density agglomeration, which also indicates that the population density of China's cities shows spatial autocorrelation, and so it needs to be interpreted by using the spatial econometric model.

(3) Spatial model selection.

This paper selects models suitable for the research topic of this paper based on the nature of different spatial econometric models. Spatial regression is used to deal with spatial effects by adding the spatial autoregressive process to the model setting, thus parameterized the spatial autocorrelation process. One of the most classical models is the Spatial Autoregressive Model (SAR).The most classical of these models are the Spatial Autoregressive Model (SAR)(As shown in equation 4) and the Spatial Error Model (SEM)(As shown in equation 5).

$$y = \rho Wy + X\beta + \varepsilon, \quad \varepsilon \sim N\left(0, \ \sigma^2 I_n\right) \qquad \text{(Eq.4)}$$

$$y = X\beta + \mu, \quad \mu = \lambda W\mu + \varepsilon, \varepsilon \sim N\left(0, \ \sigma^2 I_n\right) \qquad \text{(Eq.5)}$$

And these two are special form of Spatial Durbin Error Model (SDEM) (As shown in equation 6) and General Spatial Autocorrelation Model (SAC) (As shown in equation 7). Spatial Durbin Error Model (SDEM) and General Spatial Autocorrelation Model (SAC) are special form of the General Nesting Spatial Model (GNS)

(As shown in equation 8).

$$y = X\beta + WX\theta + \mu, \quad \mu = \lambda W_2 \mu + \varepsilon, \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.6}$$

$$y = \rho W_1 y + X\beta + \mu, \quad \mu = \lambda W_2 \mu + \varepsilon, \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.7}$$

$$y = \rho W_1 y + X\beta + WX\theta + \mu, \quad \mu = \lambda W_2 \mu + \varepsilon, \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.8}$$

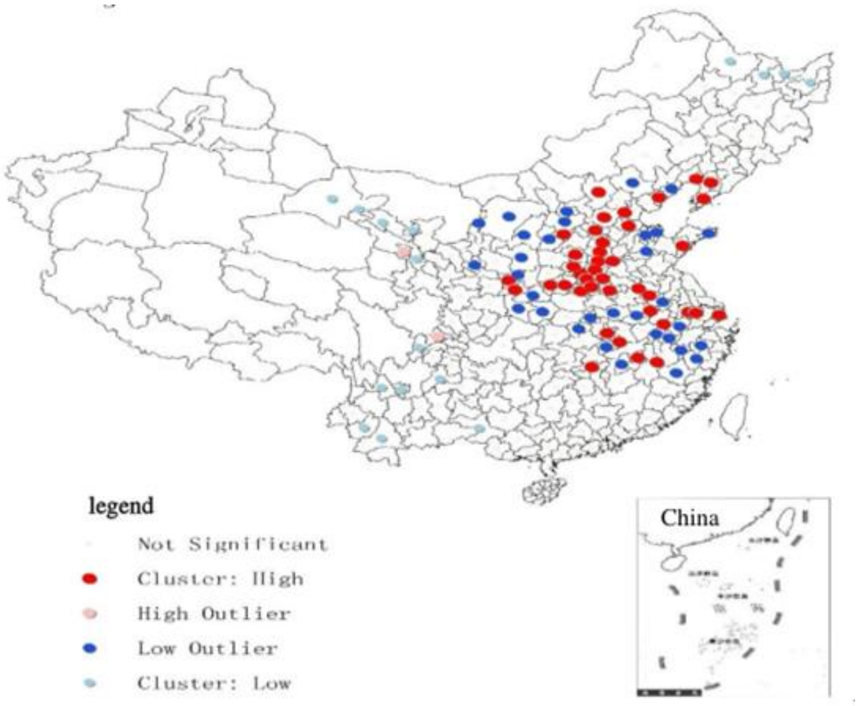

Figure 7 Localized spatial autocorrelation of population density in China's urban areas.

Other special forms of generalized nested spatial models are the Spatial Durbin Model (SDM) (As shown in equation 9). A further special form is the independent variable null similar to the classical linear model lag model（Spatial Lag Of X Model, SLX）(As shown in equation 10). If all the spatial autocorrelation coefficients are made to be 0, it degenerates into the Classical Linear Regression Model (CLR) (As shown in equation 11).

$$y = \rho Wy + X\beta + WX\theta + \varepsilon, \quad \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.9}$$

$$y = X\beta + WX\theta + \varepsilon, \quad \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.10}$$

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.11}$$

In Eqs. (4)-(11), ρ and λ are the full domain spatial autocorrelation coefficients of the explanatory variables as well as the random perturbation terms of the model. Respectively, W is the spatial weight matrix, β and θ are the regression coefficients, μ is the random perturbation term in the presence of spatial autocorrelation, and ε is the random perturbation term after eliminating spatial autocorrelation (spatial dependence).The interrelationships between the various models are shown in Figure 8.

In the above model, the Spatial Autocorrelation Model (SAC) introduces spatial effects into both the explanatory variables and the random perturbation term of the model, and in the model transformation makes the explanatory variables also have spatial utility (Eqs. 12-13). Since the three hypotheses of this paper will be included as explanatory variables in the econometric model, and they are all assumed to have spatial autoregressive effects on population agglomeration. Therefore, the SAC model is more suitable for the research topic of this paper.

$$y = \rho Wy + X\beta + \lambda W\mu + \varepsilon, \quad \varepsilon \sim N\left(0, \quad \sigma^2 I_n\right) \tag{Eq.12}$$

$$y = \left(I - \rho W\right)^{-1} X\beta + \left(I - \rho W\right)^{-1} \lambda W\mu + \left(I - \rho W\right)^{-1} \varepsilon \tag{Eq.13}$$

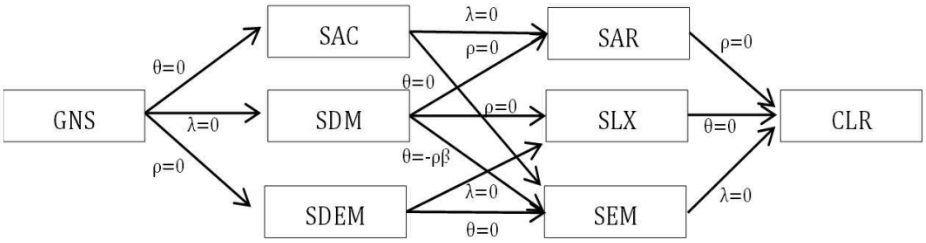

Figure 8 Spatial regression model setup.

Because Figure 4 shows that the explained variables are linearly related to each of the explanatory variables, In this paper, three types of linear models, SAC, SEM and SAR, were selected for comparative regression respectively.

The method of setting the spatial weight matrix still follows the way of setting in the spatial autocorrelation test part (part 3.2) of this paper. The threshold is whether the distance between two cities reaches $d = 1198300.76$ meters. This can maximize the spatial autocorrelation effect in the model and fully reflect the spatial autoregressive mechanism of population in the process of urban agglomeration.

(4)Endogenous problems

Since there is spatial heterogeneity and spatial autocorrelation in the data, which makes the random perturbation term of the spatial econometric model exist heteroskedastic and serially correlated, the OLS estimator is an invalid estimator. And because the power mechanism of SAC and SAR models is a spatial autoregressive process, as shown in equation (14), makes the model create an endogeneity problem, leading to OLS estimation that is biased.

$$\left(I-\rho W\right)^{-1} = I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \cdots\cdots \tag{Eq.14}$$

Therefore, the thesis uses the spatial first-order lag terms of the explanatory variables as instrumental variables (IV), and uses the Generalized Spatial 2-Stage Least Squares (GS2SLS) method for parameter estimation.

5. Estimates and Spatial Heterogeneity

Due to the existence of multicollinearity in the variable of employment opportunity (Inenter) and the variable of circulation economy size (Insale), this paper separates them for spatial regression. The results are shown in Table 4. The coefficients of Inenter, Insale and Inhospi are all positive, indicating that employment opportunities, size of circulation economy and public service capacity are positively related to population density, of which the coefficients of Inenter and Insale pass the t-test of significance level 0.05, and the coefficient of Inhospi passes the t-test of significance level 0.01, indicating that these three factors are indeed the causes of population agglomeration. The paper rejected the original hypothesis of homoscedasticity using the Breusch-Pagan heteroscedasticity test with a p-value of 0.0001. The heteroskedasticity of the random disturbance term of the model indicates that the spatial disequilibrium leading to China's megacity problem can be interpreted not only as spatial autocorrelation (spatial dependence), but also as a kind of spatial heterogeneity. In summary, all three hypotheses are valid.

Table 4 Parameter estimation results of SAR, SEM and SAC models.

Variant	**SAR**	**SEM**	**SAC**	**SAR**	**SEM**	**SAC**
C	-4.204***（0.009）	1.574***（0.004）	-4.550***（0.000）	-5.165***（0.000）	-0.009（1.000）	-5.437***（0.000）
Lnenter	0.119**（0.033）	0.131**（0.021）	0.137***（0.008）	－ － －	－ － －	－ － －
Lnsale	－ － －	－ － －	－ － －	0.195**（0.017）	0.193**（0.022）	0.198**（0.012）
Lnhospi	0.458***（0.000）	0.449***（0.000）	0.433***（0.000）	0.342***（0.003）	0.341***（0.003）	0.337***（0.003）
$\widehat{\rho}$	0.904***（0.000）	－ － －	0.977***（0.000）	0.867***（0.000）	－ － －	0.911***（0.000）
$\widehat{\lambda}$	－ － －	0.820***（0.000）	-0.940***（0.008）	－ － －	1.000***（0.000）	-0.488（0.138）
Pseudo R^2	0.382	0.344	0.381	0.372	0.337	0.373

Note: ***, **, * are the coefficients of the variables that passed the T-tests of 0.01, 0.05 and 0.1.

6. Robustness tests

6.1 Full Domain Robustness Analysis

This paper ignores the effect of spatial autocorrelation in the explanatory variables of the SAC model as well as in the random perturbation term, and uses the classical linear model and OLS estimation method to carry out the robustness test with 289 prefecture-level and above cities in 2015 as the statistical samples, and the results of parameter estimation are shown in Table 5. The symbols of the estimated parameters in Table 5 are consistent with Table 4, indicating that the estimation results in Table 4 are robust.

Table 5 OLS parameter estimation results.

Variant	OLS	t -statistic
C	1.196**	1.716
Lnenter	0.151***	2.407
Lnsale	0.124	1.293
Lnhospi	0.287***	2.375

Note: ***, **, * are the coefficients of the variables that passed the T-tests of 0.01, 0.05 and 0.1.

6.2 Regional Robustness Analysis

This paper further considers the impact of spatial sample selection on the robustness of the model.

In this paper, 100 cities are randomly selected from the sample of 289 prefecture-level and above cities in 2011, and the robustness test is conducted using OLS, and the parameter estimation results are shown in Table 6.

The symbols of the estimated parameters in Table 6 are consistent with Table 4, indicating that the estimation results in Table 4 are robust.

Table 6 Parameter estimation results for 100 prefecture-level cities.

Variant	OLS	t -statistic
C	0.597	0.441
Lnenter	0.096	0.831
Lnsale	0.283*	1.767
Lnhospi	0.148	1.061

Note: ***, **, * are the coefficients of the variables that passed the T-tests of 0.01, 0.05 and 0.1.

6.3 Spatial Cointegration Tests

Fingleton(1999)[30] was the first to extend concepts such as cointegration of time series to spatial econometrics, using Monte Carlo simulations to study pseudo spatial regression, spatial unit root and spatial cointegration. Kosfeld and Lauridsen(2006) [31] proposed a unit root test based on the LM test. Lee and Yu(2009) [32] found that the occurrence of pseudo-regression in spatial regression models is weaker compared to the time series case, but still exists. In this paper, the model is tested for robustness using the spatial unit root test proposed by Kosfeld and Lauridsen (2006) [31]. If the spatial variances of the explanatory variables are stable, i.e.

$$\Delta y = y - Wy = \left(I - W\right) y \qquad \text{(Eq.15)}$$

where $\Delta y = \left(I - W\right)$ is the spatial difference operator. The error term of the pseudo-regression model is non-smooth, and should contain a unit root, which can be expressed as Equation 15 and Equation 16.

$$y = X\beta + \mu \qquad \text{(Eq.16)}$$

$$\mu = \lambda W \mu + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I_n\right) \qquad \text{(Eq.17)}$$

where $\lambda = 1$. The original hypothesis H_0 is chosen: unsmooth $\mu = \lambda W \mu + \varepsilon$, i.e. $\mu = \Delta^{-1}\varepsilon$, which is equivalent to $\mu = \Delta^{-1}\varepsilon$, and therefore the model is equivalent to Equation 18.

$$\Delta y = \Delta X \beta + \varepsilon \tag{Eq.18}$$

Since the error term for $\Delta y = \Delta X \beta + \varepsilon$ is a white noise, this Lagrange multiplier test for the spatial error model for the equation after differencing the variables is expressed in Equation 19.

$$LM_{\lambda} = \frac{\left(ne'We / e'e\right)^2}{tr\left(\left(W + W'\right)W\right)} \tag{Eq.19}$$

where e is the residual obtained from the spatial difference equation $\Delta y = \Delta X \beta + \varepsilon$ through OLS regression. The Lagrange multiplier test statistic for the residuals obtained by $y = X\beta + \varepsilon$ through OLS regression is abbreviated as LM_{λ}. In this paper, 289 prefecture-level and above cities in 2011 are used as statistical samples, and the test results are shown in Table 7, which indicates that the Lagrange multiplier test LM_{λ} for the original equation, and the Lagrange multiplier test DLM_{λ} for the spatial difference equation, are both significantly positive, negating the original hypothesis H_0. Thus reflecting a smooth spatial autoregression with no pseudo-regression.

Table 7 Results of spatial cointegration tests.

Cointegration test	Statistic
LM_{λ}	89.904***
DLM_{λ}	34.825***

Note: ***, **, * are the coefficients of the variables that passed the T-tests of 0.01, 0.05 and 0.1.

7. Conclusion and Reflection on Countermeasures

The spatial imbalance of employment opportunities, distribution and economic layout, and public services is directly manifested in the disparity of equal employment opportunities, market prosperity, and public service capacity among different cities, and the above three factors are significantly and positively correlated with the population density, which leads to irrational migration and agglomeration of the population, and is the reason for China's megacity problem. In the national regional policies and local explorations aiming at relieving the population of big cities and promoting the synergistic development of regions, the first thing to do is to solve the problem of the unification of the above spatial imbalance in the process of China's new-type urbanization. Promoting the orderly transfer of industries between regions, relocating regional trade markets and balancing the allocation of public service resources are the paths to combating China's megacity problem and promoting economic and social integration. In promoting industrial transfer, we can learn from Shen Tiyan, Qi Zixiang et al. (2016) [33] proposed artificial intelligence algorithms, the use of economic engineering ideas, through the transfer of enterprises to be transferred out of Beijing and Hebei to take over the development zone of the bilateral matching of market design, the evacuation of non-capital functions.

Acknowledgment

In this article, we sincerely thank and appreciate all the senior managers and experts who helped us during the implementation process of this research, especially in the feasibility section of the proposed models.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Paelinck P. L’Efficacité Des Mesures De Politique éConomique RéGionale [A]. L’ Efficacité Des Mesures De Politique éConomique RéGionale [M]. CERVNA, 1967.
2. Moran P. The Interpretation of Statistical Maps [J]. Journal of the Royal Statistical Society, 1947, 10(2):

243-251.

3. Geary R. C. The Contiguity Ratio and Statistical Mapping [J]. The Incorporated Statistician, the Incorporated Statistician 1954,5(3):115-145.
4. Cliff A., Ord J.K. Testing for Spatial Autocorrelation Among Regression Residuals [J]. Geographical Analysis, 1972,(4):267-284.
5. Ord J.K. Estimation Methods for Models of Spatial Interaction [J]. Journal of the American Statistical Association, 1975,70(349): 120-126.
6. Paelinck P., Klaassen L.H. Spatial Econometrics [M]. Saxon House,1979.
7. Anselin L. Spatial Econometrics: Methods and Models [M].Kluwer Academic Publishers, 1988.
8. Anselin L. Spatial Econometrics [A]. Palgrave Handbook of Econometrics [M]. Palgrave Macmillan, 2006.
9. Pace R.K., Barry R. Quick Computation of Spatial Autoregressive Estimators [J]. Geographical Analysis, 1997, 29(3):232-246.
10. Barry R., Pace R.K. LA Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices [J]. Linear Algebra and Its Applications, 1999, 289(1): 41-54.
11. Pace R.K., Lesage J.P. Chebyshev. Approximation of Log-Determinants of Spatial Weights Matrices [J]. Computational Statistics and Data Analysis 2004, 45(2): 179-196.
12. Zhang Y., Leithead W.E. Approximate Implementation of Logarithm of Matrix Determinant in Gaussian Processes [J]. Journal of Statistical Computation and Simulation, 2007, 77(4): 329-348.
13. Bivand R., Hauke J., Kossowski T. Computing the Jacobian in Gaussian Spatial Autoregressive Models, an Illustrated Comparison of Available Methods [J]. Geographical Analysis 2013,45(2):150-179.
14. Anselin L. Estimation Methods for Spatial Autoregressive Structures [D]. Cornell University, 1980.
15. Kelejian H.H., Prucha I.R. A Generalized Spatial Two Stage Least Squ ares Procedures for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances [J]. Journal of Real Estate Finance and Economics 1998, 17(1): 99-121.
16. Kelejian H.H., Prucha I.R. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model [J]. International Economic Review 1999, 40(2): 509-533.
17. Kelejian H.H., Prucha I.R. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances [J]. Journal of Econometrics 2010,157(1): 53-67.
18. Drukker D., Egger P., Prucha I. On Two-Step Estimation of a Spatial Autoregressive Model with Autoregressive Disturbancesand Endogenous Regressors [J]. Econometric Reviews 2013, 32(5):686-733.
19. Hepple L. Bayesian Analysis of the Linear Model with Spatial Dependence [A]. Exploratory and Explanatory Statistical Analysis of Spatial Data [M]. Martinus Nijhoff, 1979.
20. Lesage J.P. Bayesian Estimation of Spatial Autoregressive Models [J]. International Regional Science Review 1997,20(2):113-129.
21. Wang X., Kockelman K.M., Lemp J.D. The Dynamic Spatial Multinomial Probit Model, Analysis of Land Use Change Using Parcel-Level Data [J]. Journal of Transport Geography, 2012,24(4):77-88.
22. Getis A. Spatial Filtering in a Regression Framework, Examples Using Data on Urban Crime, Regional Inequality, and Government Expenditures [A]. New Directions in Spatial Econometrics [M]. Springer-Verlag, 1995.
23. Anselin L., Kelejian H.H., Testing for Spatial Error Autocorrelation in the Presence of Endogenous Regressors [J].International Regional Science Review, 1997, 20(1): 153-182.
24. Baltagi B.H., Li D. LM Tests for Functional Form and Spatial Error Correlation [J]. International Regional Science Review,2001, 24(2): 194-225.
25. Anselin L. Spatial Econometrics [A]. A Companion to Theoretical Econometrics [M]. Blackwell, 2001. Spatial Regression, Spatial.

26. Lauridsen J., Kosfeld R. A Test Strategy for Spurious Nonstationarity, and Spatial Cointegration [J]. Papers in Regional Science, 2006,85(3):363-377.
27. Kelejian H.H. A Spatial J-Test for Model Specification Against a Single or a Set of Non-Nested Alternatives [J]. Letters in Spatial and Resource Sciences, 2008, 1(1): 3-11.
28. Amaral P.V., Anselin L. Finite Sample Properties of Moran's I Test for Spatial Autocorrelation in Tobit Models [J]. Papers in Regional Science, 2014, 93(4): 773-781.
29. Anselin L., Florax R. J. G. M. New direction in spatial econometrics [M]. Springer-Verlag Berlin and Heidelberg GmbH& Co. K, 1995.
30. Fingleton B. Spurious spatial regression: Some Monte Carlo results with spatial unit root and spatial cointegration [J]. Journal of Regional Science, 1999, 39: 1-19.
31. Lauridsen J., Kosfeld R. A test strategy for spurious spatial regression, spatial nonstationarity, and spatial cointegration [J]. Papers in Regional Science, 2006, 85: 363-377.
32. Lee L. F., Yu J. H. Spatial nonstationarity and spurious regression: The case with a row-normalized spatial weights matrix[J]. Spatial Economic Analysis, 2009, 4: 301-327.
33. Shen Tiyan, Qi Zixiang, Wang Yanbo. Market design for orderly inter-regional transfer of industries in Beijing-Tianjin-Hebei － Based on bilateral matching algorithm[J]. Economist, 2016(4):42-52.