A Spatial Analysis of Educational Inequality in Mainland China

April 16th, 2012

Stats 499 Paper

Zhengyao Wang

Instructor: Professor Ed Rothman

Introduction

As a developing country, China has a relatively high degree of inequality in many aspects due to limited resources. Education, as an important factor closely associated with economic development and people's welfare, is raising increasing concern, and it deserves our study. Nowadays, spatial analysis has become an important tool in studying social science data with geographical features. It helps people identify spatial patterns and distribute resources more effectively. The purpose of this thesis is to perform a spatial analysis of education in mainland China, and hopefully to shed some light on the geographic pattern of education in China. Previously, similar issues in other countries, such as Bangladesh, have been discussed (see Zahiduzzaman, Quasem, Khan, & Rahman, 2010). In this thesis, some other spatial analysis techniques, such as geographically weighted regression, are applied to analyze the inequality problem. Special administrative regions, such as Hong Kong and Macau, are excluded. Centrally administered municipalities, such as Beijing and Shanghai, and autonomous regions, such as Xinjiang and Tibet, are included in the discussion. All are referred to as "provinces" in this paper because they are at the same level as a province.

General Pattern

Nine-year compulsory education is a basic national policy in China. China's fundamental education level has been improved greatly in the last decade. According to the National Bureau of Statistics of China, the national illiteracy rate has decreased from 6.72% in 2000 to 4.08% in 2010. However, disparity in fundamental education among provinces remains an important issue in China.

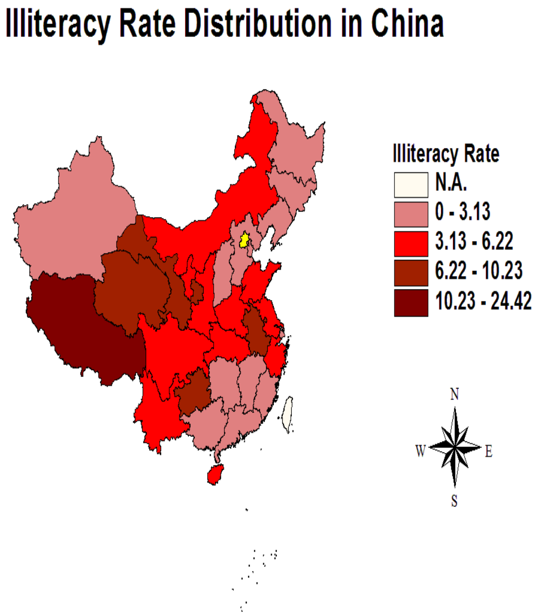


Figure 1. Illiteracy Rate Distribution in China[1]

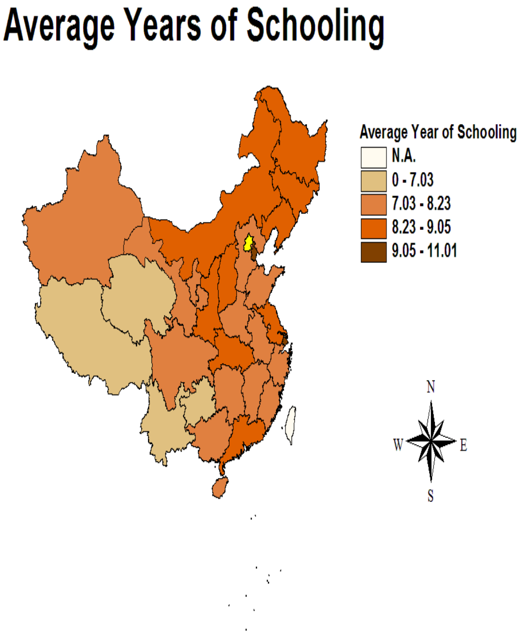


Figure 2. Average Years of Schooling in China[2]

Figure 1 demonstrates that spatial clustering of illiteracy rates clearly exists in mainland China. Provinces geographically close to each other tend to share similar illiteracy rates. The illiteracy rate is considerably lower in northeastern China and southeastern China, while the rate remains high in the southwestern provinces. Among all the provinces, Tibet is highlighted due to its uncommonly high illiteracy rate of 24.42%. Not surprisingly, the areas suffering from high illiteracy rates are also the most undeveloped regions in China. Figure 2 displays the average

[1] Data Source: National Bureau of Statistics of China, *China Statistical Year Book 2011*. The software used to draw the map is ArcView GIS. GIS data of China are retrieved from http://china-archive.library.tamu.edu/datasets/chinaarchive/Geospatial/General%20GIS%20Data%20Sets/Data/ (The China Archive at Texas A&M University). Chongqing is not included in the shape file.

[2] The average year of schooling is calculated using data from the National Bureau of Statistics of China, "Comparison of Population with Various Education Attainment per 100000 Persons by Region" in *China Statistical Year Book 2011*. The year of primary school attainment counts as 6 years, junior secondary school counts as 9 years, senior second school counts as 12 years, junior college and above counts as 16 years.

years of schooling in each province. Similarly, provinces in southwestern China have the lowest average years of schooling. According to the moment condition suggested by Anselin (1999), we notice that $Cov[l_i\ ,\ l_j] \neq 0$ for $i \neq j$, where l_i and l_j are the illiteracy rates for the *ith* and *jth* observations.

Methodology

Moran's I (Moran, 1950) is a popular statistic with which to study the global spatial autocorrelation of the variables people are interested in. In addition to the global spatial autocorrelation measurement, local indicators of spatial correlation (LISA) have been suggested to further analyze local spatial patterns (Anselin, 1993). Moreover, in exploring the relationship among different variables, geographically weighted regression (Fotheringham, Charlton, & Brunsdon) has received attention due to its recognition of the spatial pattern of the relationship. It allows researchers to examine the relationship among explanatory variables and dependent variables for each region, while making full use of the spatial information contained in these data. These techniques will be applied in this thesis to study the educational pattern of China. However, a certain weakness of the spatial analysis methods should be kept in mind. Although these methods pay close attention to the geographic proximity, they may fail to consider the "strength of links" among regions, such as the condition of transportation among provinces in our case (Cliff & Ord, 1981).

Moran' I and Moran Scatterplot

Moran's I (Moran, 1950) allows us to study the spatial autocorrelation of a variable we are interested in. In computing Moran's I, we use the binary contiguity matrix *W*; namely,

$w_{ij} = 1$ when the *ith* observation shares boarder with the *jth* observation and $w_{ij} = 0$ otherwise.[3] In addition, an observation does not count as a neighbor of itself. The matrix is row standardized and *N* is the number of observations.

Moran's I for illiteracy rate:

$$I_l = \frac{N}{\sum_i \sum_j w_{ij}} * \frac{\sum_i \sum_j w_{ij}(l_i - \bar{l})(l_j - \bar{l})}{\sum_i (l_i - \bar{l})^2} = 0.1603$$

where l_i is the illiteracy rate for *ith* observation and $\bar{l}$ is the mean of illiteracy rate.

A positive Moran's I value indicates the existence of positive spatial autocorrelation. Under the assumption of no spatial autocorrelation,

$$E[I] = -\frac{1}{(N-1)} = -0.033$$ (Moran, 1950)

Hence, 0.1603 indicates that provinces geographically closer to each other tend to have moderately similar illiteracy rates. However, the Moran's I statistic is lower than what we would expect from Figure 1, which suggests a more significant positive spatial autocorrelation. A Moran scatterplot is useful for diagnosis of spatial autocorrelation (Anselin & Bao, 1997). The normalized observation is plotted against its own normalized spatial lag, where spatial lag refers to the average of its neighbors' values. The spatial lag of illiteracy rate is $W * z$, where W denotes the contiguity matrix discussed above and z is a normalized vector of illiteracy rates in each province.

[3] Although there is discontiguity between Hainan Province and Guangdong Province, their weight is assigned to be 1 for analysis because Guangdong is the nearest province of Hainan.

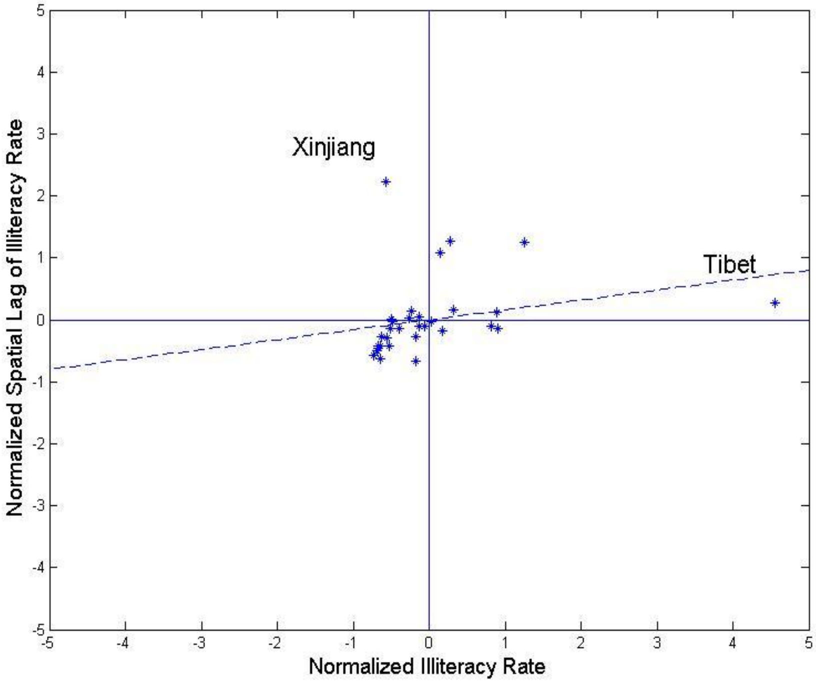


Figure 3: Moran Scatterplot of Illiteracy Rate

Two observations in the Moran scatterplot merit our attention. The top left point corresponds to Xinjiang; despite the high illiteracy rate of its neighbor provinces, Xinjiang itself has an illiteracy rate lower than the national level. The right point corresponds to Tibet, which has illiteracy rate four standard deviations above the mean, although the average of Tibet's neighbors is only slightly above the mean of illiteracy rate. Following Anselin and Bao (1997), the Moran's I corresponds to the slope coefficient of regressing $W * z$ on z. Tibet drags down the slope as a high leverage point and extreme outlier. Moran's I will increase to 0.3715 if we exclude Tibet and Xinjiang, which indicates a stronger positive spatial autocorrelation in the rest of China.

Local Indicator of Spatial Autocorrelation

The traditional Moran's I statistic provides us with a measurement of the global spatial autocorrelation of illiteracy rates; however, it fails to capture the local spatial pattern. Local indicators of spatial autocorrelation, such as local Moran's I, have been suggested to compensate for the local measurement (Anselin, 1993). The local Moran's I statistic could be calculated as below:

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij}\, z_j \qquad \text{(Anselin, 1993)}$$

w_{ij} is defined in the same way as discussed in global Moran's I, z_i is the normalized observation of illiteracy rates, and $m_2 = \sum_i \frac{{z_i}^2}{n}$. Anselin and Bao (1997) further suggested that plotting the LISA significance map contributes to identifying the local clustering.

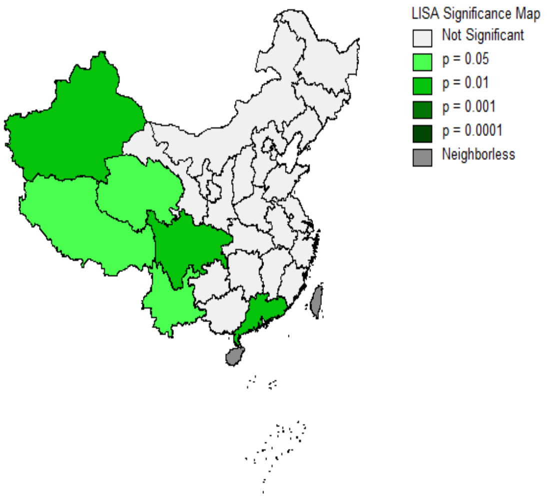


Figure 4: LISA Significance Map[4]

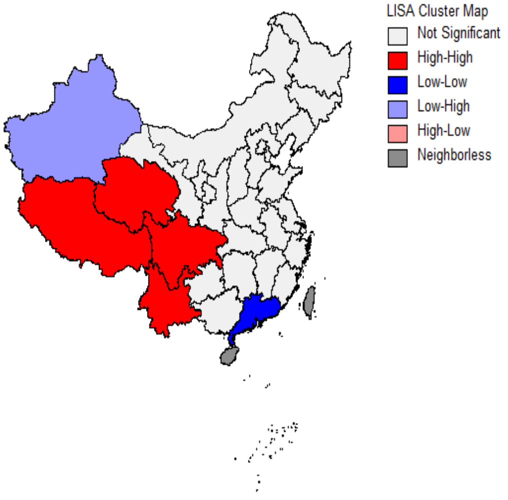


Figure 5: LISA Cluster Map[5]

[4] [5]. The LISA significance and cluster maps are created by the software OpenGeoDa provided by GeoDa Center at Arizona State University. The software is downloaded from http://geodacenter.asu.edu/software/downloads

Among those regions with significant local indicators of spatial autocorrelation, Xinjiang located at the northwestern corner of China, has a negative local Moran value, indicating a significant negative local spatial autocorrelation. Xinjiang has a surprisingly low illiteracy rate of 2.36% while its neighbors all suffer from high illiteracy rates. The provinces in the southwest of China, including Yunnan, Sichuan, Tibet and Qinghai all have significant positive local Moran's I values. As we notice on the map, high illiteracy levels are consistent and persistent in the southwestern part of China. China's government should devote more effort in ensuring the fundamental education in those provinces.

Geographically Weighted Regression

A traditional regression model, such as the linear regression model, will return the global relationship among explanatory variables and dependent variables. However, it is possible that the relationship would vary across the country; it is sloppy to assume such a relationship would remain constant among different regions. Geographically weighted regression has been suggested to deal with this issue (Fotheringham, Charlton, & Brunsdon).

Traditional Linear Regression Model: $y_i = \alpha_0 + \sum_k \alpha_k x_{ik} + \varepsilon_i$ (1)

Geographically Weighted Regression Model: $y_i = \alpha_{i0} + \sum_k \alpha_{ik} x_{ik} + \varepsilon_i$ (2)

Moreover, the estimator of $a_i$is given by $\widehat{a_i} = (x^t W_i x)^{-1} x^t W_i y$ (Fotheringham, Charlton, & Brunsdon).

$$\text{where} \quad W_i = \begin{pmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{pmatrix}$$

One choice of w_{ij} is suggested by Fotheringham, Charlton and Brunsdon by setting $w_{ij} = \exp(-\beta {d_{ij}}^2)$, where β denotes a distance-decay parameter, and d_{ij} is the distance between *ith* observation and *jth* observation. They further suggested some optimizing criteria of the distance-decay parameter β using cross-validation method and Golden Section search approach, but this will not be addressed here. Several choices of β and their effects are displayed below.

Table 1. Weights under Different Choices of Distance Decay Parameter [6]

	Distance	Weight under $\beta = 0$	Weight under $\beta = \frac{1}{1500000}$	Weight under $\beta = \frac{1}{3000000}$	Weight under $\beta = \frac{1}{4500000}$
Beijing – Tianjin	104 km	1	0.9928	0.9964	0.9976
Beijing – Shanghai	1065 km	1	0.4695	0.6852	0.7772
Beijing – Guangdong	1889 km	1	0.0927	0.3044	0.4525
Beijing – Xinjiang	2417 km	1	0.0204	0.1427	0.2730

From Table 1 above let us observe that the further apart two observations are located, the smaller their mutual effects are. The distance between Beijing and Shanghai is around 1065 kilometers, and Shanghai accounts for only 68.52% of the weight under $\beta = \frac{1}{3000000}$ when we estimate the model for Beijing; and the weight of Xinjiang is almost negligible due to the long distance. Different from Equation (1), Equation (2) allows each observation to have its unique parameter of each explanatory variable to describe the underlying relationship within that area

[6] The distance between two provinces is measured as the distance between their capital cities. Data is retrieved from http://wenku.baidu.com/view/a1bbae8583d049649b6658af.html

and its neighbors. By studying the spatial pattern of these parameters, we can possibly find out how the effects of explanatory variables vary across the country.

Application

I use the geographically weighted regression to explore how effectively the educational funds have been used in each province. Consider the following model:

$$y_i = \alpha_0 + \alpha_1 * x_i + \varepsilon_i$$

where y_i denotes the average year of schooling in *ith* observation, and x_i corresponds to the educational fund per capita in *ith* observation. α_0 is the constant, α_1 is the parameter of educational funds per capita, and ε_i is the error term.

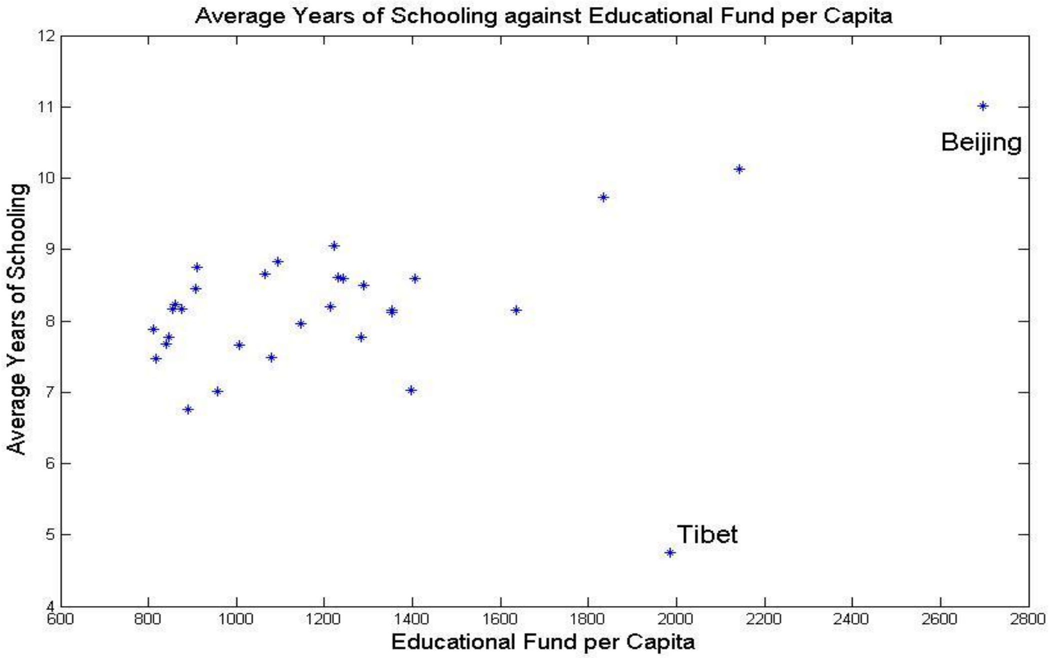


Figure 6. Educational Funds Per Capita against Average Years of Education for each Province[7]

[7] The educational funds data and population data are from *China Statistical Year Book 2011*.

Figure 3 shows a fairly strong linear relationship between educational fund per capita and the average year of schooling, with Tibet as an exception. Tibet suffers from poor education although it has high educational fund support. Tibet is one of those autonomous regions with unique characteristics, such as religion and ethnicity. The majority of its population adheres to Tibetan Buddhism, which greatly differs from the rest of mainland China. In addition, China's government implements special beneficiary policies toward Tibet; hence, we drop it out of the regression model and discuss it separately.

Using ordinary least squared method, $\hat{\alpha} = (X^T * \mathrm{X})^{-1} * X^T * y$, the national model therefore is estimated as:

$$y_i = 6.37850 + 0.00158 * x_i + \varepsilon_i$$

t-statistic (18.83) (5.95) $R^2 = 0.56$

The linear regression model above suggests that on average 1000 Yuan increase in educational fund per capita is associated with an improvement of 1.58 years in average years of schooling. Then we apply geographically weighted regression to estimate the model for each region. The model is described as:

$$y_i = \alpha_{i0} + \alpha_{i1} * x_i + \varepsilon_i$$

The estimators of constant term and parameters of educational fund per capita for each observation are given as:

$$\widehat{\alpha_i} = (X^T W_i X)^{-1} X^T W_i y$$

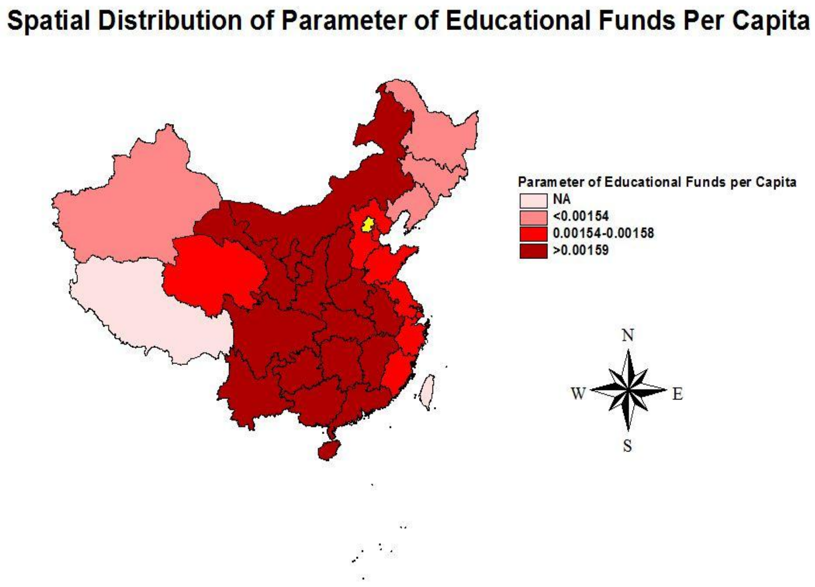

Spatial Distribution of Parameter of Educational Funds Per Capita
Parameter of Educational Funds per Capita
NA
<0.00154
0.00154-0.00158
>0.00159
N
W
E
S

Figure 7. Spatial Distribution of the Educational Funds per Capita Parameter under $\boldsymbol{\beta = \frac{1}{3000000}}$

Figure 7 suggests that the effects of educational funds vary across the country. As we notice, the parameters in the western region and northeastern region tend to be smaller than the other parts of China. It implies the educational funds there may have not been spent as effectively as their counterparts. Table 2 in the appendix displays the estimates for each province under various choices of distance decay parameter β.

Conclusion

Positive spatial autocorrelation of education is relatively clear in most parts of China. Provinces geographically close to each other tend to have similar educational levels. Generally speaking, provinces in western China are less developed and they suffer from poor education as well. China’s government needs to continue supporting fundamental education in less developed regions in order to get rid of the stubbornness of illiteracy there.

As we may expect, the effect of educational funds varies across the country. The effects of educational investment are less prominent in the western and northeastern China than the other part of the country. Furthermore, although Tibet enjoys high educational fund per capita, its educational level is considerably lower than the other parts of the country. A possible explanation is that infrastructures there are not as advanced as in the rest of mainland China. Other factors, such as different ethnicities and religions, could also contribute to the difference discussed above.

Limitation and Further Study

The data from the National Bureau of Statistics of China are highly aggregated. When a local government implements regional educational policy, it is important to recognize the large inequality within each province, particularly the disparity between urban and rural regions. By doing a spatial analysis, local government could more effectively distribute its limited resources, and hence improve the education level in each region. In addition, we drop Tibet out of the regression model due to its unique characteristics; in the future, it is worthwhile to study why and how much the cultural and ethnic differences contribute to the discrepancy in the education levels. China is a country with 56 ethnic groups and ethnic conflicts occasionally arise partly because of inequality; therefore, understanding the differences among ethnic groups is essential to maintaining social stability.

Thanks to the rapid development in spatial statistics, more sophiscated statistical methods and software make a deeper and more complete spatial analysis possible. Hopefully spatial analysis could play a more important role in China's policy making process in the near future.

Acknowledgement

I deeply appreciate the help of Professor Ed Rothman of the Statistics Department at University of Michigan, especially for leading me into the field of spatial analysis.

I would also like to thank Professor Jeff Smith of the Economics Department, Professor Ben Hansen of the Statistics Department, and Ray McDaniel at the Sweetland Center for Writing at the University of Michigan for providing very helpful suggestions.

Last but not least, I would like to thank Texas A&M University for making the China GIS data available. I also thank Arizona State University GeoDa Center for making OpenGeoDa software available.

References

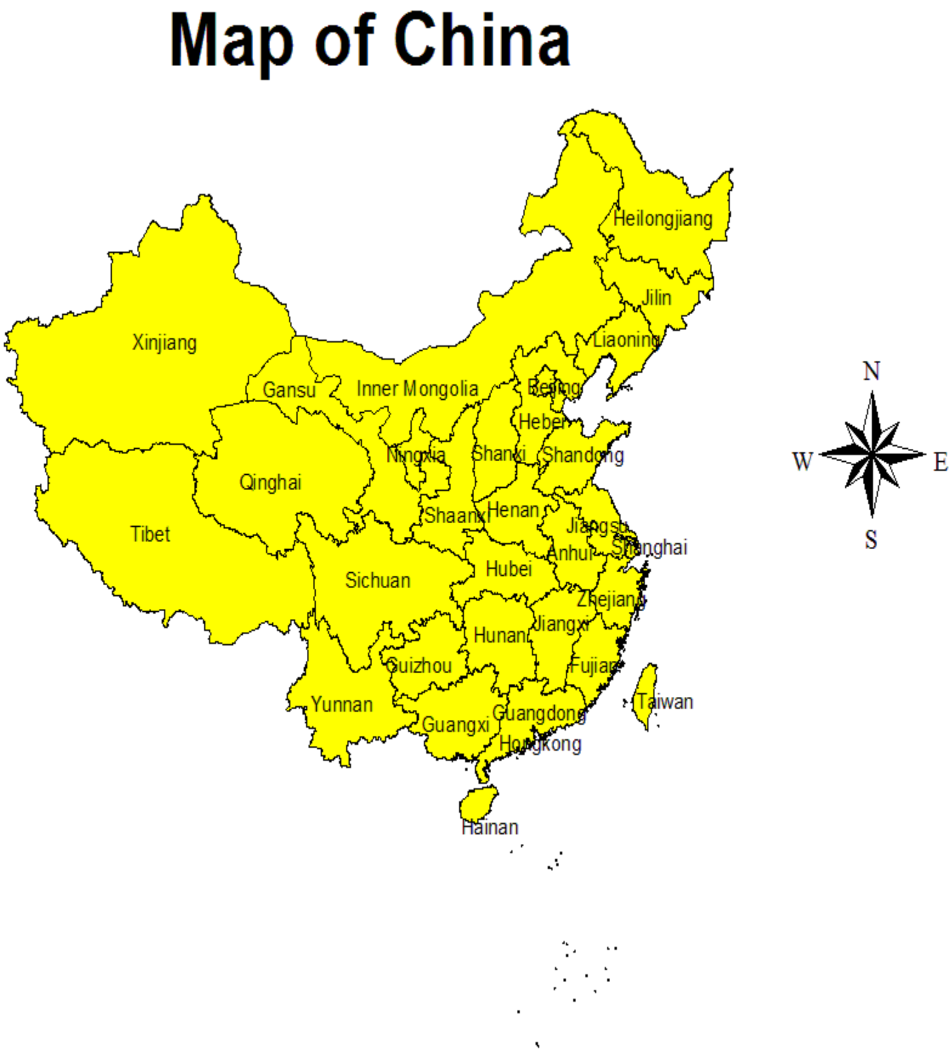
1. Anselin, L. (1993). Local Indicators of Spatial Autocorrelation – LISA. Regional Research Institute Research Paper 9331.
2. Anselin, L. (1999). Spatial Econometrics. Bruton Center, School of Social Science, University of Texas at Dallas, Richardson, TX.
3. Anselin, L., & Bao, S. (1997). Exploratory Spatial Data Analysis Linking SpaceStat and ArcView. Chapter 3 in *Recent Developments in Spatial Analysis, Spatial Statistics, Behavioral Modeling and Computational Intelligence* M.M. Fischer and A. Getis (Eds.). Berlin: Springer-Verlag: 35-59.
4. Cliff, A.D., & Ord, J.K. (1981). *Spatial Processes Models & Applications*. London: Pion.
5. Fotheringham, A.S., Charlton, M., & Brunsdon, C. Measuring Spatial Variations in Relationships with Geographically Weighted Regression, Chapter 4 in *Recent Developments in Spatial Analysis, Spatial Statistics, Behavioral Modeling and Computational Intelligence* M.M. Fischer and A. Getis (Eds.). Berlin: Springer-Verlag: 60-82.
6. Moran, P.A.P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika,* 37, 17-23.
7. National Bureau of Statistics of China (2011). *China Statistical Year Book 2011*. Retrieved from China Data Center at the University of Michigan, http://chinadataonline.org/member/yearbooknew/yearbook/Aayearbook.aspx?ybcode=F88F0ED0C6279E70148150AA6B635573&key=en
8. Zahiduzzaman, A.K.M., Quasem, M.N., Khan, M., & Rahman, R.M. (2010). Spatial Data Mining on Literacy Rates and Educational Establishments in Bangladesh. Proceedings of 13th International Conference on Computer and Information Technology (ICCIT 2010).

Appendix

Table 2. Estimators of Constant and Parameter of Educational Fund per Capital under Different Distance Decaying Parameters

	$\beta = \frac{1}{1500000}$		$\beta = \frac{1}{3000000}$		$\beta = \frac{1}{4500000}$		$\beta = \frac{1}{6000000}$	
Region	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$
Beijing	6.588	0.00157	6.507	0.00157	6.470	0.00157	6.449	0.00157
Tianjing	6.590	0.00157	6.508	0.00157	6.470	0.00157	6.449	0.00157
Hebei	6.499	0.00158	6.449	0.00158	6.427	0.00158	6.416	0.00158
Shanxi	6.451	0.00159	6.418	0.00159	6.405	0.00158	6.399	0.00158
Inner Mongolia	6.439	0.00161	6.429	0.00159	6.417	0.00158	6.409	0.00158
Liaoning	6.851	**0.00149**	6.661	**0.00153**	6.577	**0.00154**	6.531	**0.00155**
Jilin	6.989	**0.00144**	6.736	**0.00150**	6.627	**0.00153**	6.568	**0.00154**
Heilongjiang	7.115	**0.00139**	6.817	**0.00147**	6.681	**0.00151**	6.607	**0.00153**
Shanghai	6.534	0.00155	6.473	0.00157	6.444	0.00158	6.428	0.00158
Jiangsu	6.502	0.00156	6.443	0.00158	6.420	0.00158	6.409	0.00159
Zhejiang	6.504	0.00155	6.443	0.00158	6.420	0.00158	6.409	0.00159
Anhui	6.476	0.00157	6.419	0.00159	6.402	0.00159	6.395	0.00159
Fujian	6.471	0.00152	6.397	0.00158	6.380	0.00159	6.375	0.00159
Jiangxi	6.422	0.00156	6.346	0.00161	6.329	0.00161	6.327	0.00161
Shandong	6.533	0.00157	6.470	0.00158	6.442	0.00158	6.427	0.00158
Henan	6.438	0.00159	6.399	0.00159	6.389	0.00159	6.385	0.00159
Hubei	6.414	0.00157	6.367	0.00159	6.361	0.00160	6.361	0.00159
Hunan	6.375	0.00156	6.330	0.00160	6.330	0.00160	6.335	0.00160
Guangdong	6.359	0.00153	6.303	0.00159	6.302	0.00160	6.310	0.00160
Guangxi	6.227	0.00157	6.239	0.00160	6.255	0.00161	6.271	0.00161
Hainan	6.267	0.00155	6.252	0.00160	6.261	0.00161	6.274	0.00161
Chongqing	6.265	0.00155	6.261	0.00160	6.279	0.00160	6.294	0.00160
Sichuan	6.250	0.00152	6.243	0.00159	6.265	0.00160	6.284	0.00160
Guizhou	6.234	0.00156	6.243	0.00160	6.262	0.00161	6.278	0.00161
Yunnan	6.145	0.00155	6.193	0.00159	6.221	0.00161	6.244	0.00161
Shaanxi	6.339	0.00158	6.328	0.00160	6.335	0.00160	6.342	0.00159
Gansu	6.279	0.00153	6.284	0.00159	6.303	0.00159	6.317	0.00159
Qinghai	6.267	0.00149	6.267	0.00158	6.291	0.00159	6.307	0.00159
Ningxia	6.320	0.00157	6.335	0.00159	6.345	0.00159	6.352	0.00159
Xinjiang	6.276	**0.00130**	6.218	**0.00149**	6.254	**0.00154**	6.280	**0.00156**

Figure 8. Map of China


Map of China
Heilongjiang
Jilin
Liaoning
Xinjiang
Gansu
Inner Mongolia
Beijing
Hebei
Ningxia
Shanxi
Shandong
Qinghai
Tibet
Shaanxi
Henan
Jiangsu
Anhui
Shanghai
Sichuan
Hubei
Zhejiang
Jiangxi
Hunan
Guizhou
Fujian
Taiwan
Yunnan
Guangxi
Guangdong
Hongkong
Hainan
N
W
E
S