A METHOD USING ESDA TO ANALYZE THE SPATIAL DISTRIBUTION PATTERNS OF CULTURAL RESOURCE

Dongying ZHANG [a*], Xiajun MAO [a], Lingkui MENG [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
yineast@mail.whu.edu.cn, xiajun.mao@gmail.com, lkmeng@whu.edu.cn

Commission VI, WG VI/4

KEY WORDS: Spatial Statistics, Clustered Qualification, ESDA, Spatial Distribution Pattern, Spatial Association, Spatial Autocorrelation, Cultural Geography

ABSTRACT:

The spatial distribution pattern of cultural resource generally manifests clustered qualification to some extent, and we can infer the vicissitudes of the correlative culture by analyzing the changing rule of the spatial distribution pattern of cultural resource. However, the majority of the concerned researches are still at the level of qualitative statistics and thematic map visualization of some general features at present, without taking the characteristic into account. Aiming at making up the shortage, the paper therefore analyzes the spatial distribution of a special type of cultural resource - the ancient Toponym which is defined as the name of the places where the immigrants settled down by using Exploratory Spatial Data Analysis (ESDA) methods. We first made both qualitative analysis and quantitative analysis to confirm the existence of the clustered qualification by computing Global Moran's I and Global Geary's C values, later we made a further study about the similarity of different clustered regions and told them apart. At last, we came to a conclusion that not only the spatial distribution pattern of cultural resource is usually clustered, but also the similarity degree of the clustered regions differs from one another. In addition to revealing the spatial patterns of the Toponym distribution, this paper promotes an explicitly spatial view that has certain methodological implications for the application of spatial statistic methods in Cultural Geography research.

1. INTRODUCTION

In order to increase the scale of the population and develop the local economy in Sichuan Region, the government of Qing Dynasty adopted emigration policy which lasted more than 100 years, it's the well-known legend called "HuGuang Tian Sichuan" policy in the history, which not only affected the population distribution of China strongly, but also promoted the amalgamation of emigration cultural and local cultural directly.

The Toponym has accompanied the history in a long run, it's the cultural patrimony which not only reflect local geographical environment, but track the emigration of the ancient race and their war，it looks like a live fossil that can play the role of the indicator of culture (Zhengxiang Cheng 1992). The Toponym is the visual cultural sight that can not only reflect the distribution of emigration directly, but can be mined for plenty of cultural content (Shangji Situ 1983).

As a special type of cultural resource, the Toponym plays a key role in the research of historical geography, cultural geography and Geographical information science, and its importance has been recently recognized. For example, some historical geographers have statistically analyzed the quantity of the Toponym in every administrative unit during Qing Dynasty in Sichuan Region so as to estimate each part's immigrants and the percentage by their descent, ignoring the spatial interaction between the Toponym. With spatial techniques, GIS experts have already visualized the spatial distribution of the Toponym by using spatial interpolation and obtained clustered areas where people say the same dialect using point-based cluster analysis methods (Fahui Wang 2009). Recent researches have proved the spatial distribution patterns of cultural resources qualitatively. However, we need spatial distribution patterns expressed quantitatively as the spatial correlation ratio to support the economic and cultural researches, the higher ratio the more developed economic and smaller cultural difference. This article takes the ancient Toponym quantity into research, with ESDA to analyze the spatial distribution patterns of cultural resource so as to promote the application of spatial analysis methods in the research of Cultural Geography.

2. MAJOR ANALYSIS ISSUES AND ANALYSIS MEANS

2.1 Major Analysis Issues

Exploratory spatial data analysis (ESDA) has its origins in exploratory data analysis (EDA) which is a term coined by the American statistician John Tukey in the 1970s to describe statistical procedures used by applied statisticians when they

* Corresponding author.

were in the first stages of analyzing a new set of data. ESDA can be considered the extension of EDA methods to spatial data. Often, ESDA is used to identify data properties for four purposes as follows: detecting spatial patterns in data, detecting possible data errors ('outliers' and 'spatial outliers'), formulating hypotheses based on the geography of the data and assessing spatial models. According the first law of Geography: "Everything is related, but things nearby are more related than things far away", we know that spatial association is inherent in geographic data, when working on spatial data, analyses based on regular statistics are very likely to be misleading or incorrect, There is positive spatial association when high or low values of a random variable tend to cluster in space and there is negative spatial association when geographical areas tend to be surrounded by neighbors with very dissimilar values, all of which consists of spatial patterns, and therefore the main aim of ESDA is for patterns detection. Spatial patterns include three types: collected pattern, randomized pattern and dispersed pattern.

In the analysis of the spatial distribution of the Toponym, the major concerns are to reveal spatial patterns. The distribution of the Toponym is intrinsically spatial and, moreover, space-dependent due to the potential interactions of the long term movement of emigration. The spatial distribution pattern of the Toponym may be a reflection of the vicissitudes of the local culture and the amalgamation of emigration culture and local culture.

2.2 Major Analysis Means

Spatial analysis and techniques for measuring spatial association have been proposed in the literature. Getis and Ord family of Gi(d) statistics (Getis and Ord 1993; Ord and Getis, 1995) and Anselin's LISA(Local Indicators of Spatial Association) (Anselin 1995) are two basic local statistics of spatial association. Compared to LISA, Gi(d) statistics is more simple in detecting places with unusual concentrations of high or low values (i.e., 'hot' or 'cold' spots). On the other hand, techniques for spatial heterogeneity include the expansion method (Casetti 1972; Jones and Casetti 1992), the method of spatial adaptive filtering (Foster and Gorr1986; Gorr and Olligschlaeger 1994), the random coefficients model (Aitken 1996), the multilevel modelling (Goldstein 1987), the moving window approach (Fotheringham et al. 1997) and geographically weighted regression (GWR) (Brunsdon et al. 1996; Fotheringham et al. 1997). However, GWR is relatively a simple but effective technique for exploring spatial heterogeneity which allows different relationships existing.

In this article, we mainly use global Moran and Geary, local G-Statistics and LISA methods to analysis the degree of spatial autocorrelation of the Toponym's attributes data from global regions to local counties.

3. A BRIEF REVIEW OF GISA AND LISA

3.1 GISA

GISA is short for Global Indicators of Spatial Association, it mainly includes Moran's I indicator and Geary's C rate.

3.1.1 Moran's I

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

where x_i, x_j =the observed value at location (i, j)

$\bar{x}$ = the average of the $\{x\}$ over the n locations

w_{ij} = the spatial weight measure defined as 1 if location i and j are adjacent, or else as 0

i = contiguous to location j and 0 otherwise

The expected value and variance of the Moran I for samples of size n could be calculated according to the assumed pattern of the spatial data distribution (Cliff and Ord 1981, Goodchild 1986).

For the assumption of a normally distribution:

$$E_R(I) = \frac{-1}{n-1} \quad (2)$$

$$V_1 = \frac{n\left[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2\right]}{(n-1)(n-2)(n-3)W^2} \quad (3)$$

$$V_2 = \frac{k\left[(n^2 - n)S_1 - nS_2 + 3W^2\right]}{(n-1)(n-2)(n-3)W^2} \quad (4)$$

$$\mathrm{var}_R(I) = V_1 - V_2 - [E_R(I)]^2 \quad (5)$$

$$z_{Moran} = \frac{I - E_R(I)}{\sqrt{\mathrm{var}(I)}} \quad (6)$$

where $W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ (7)

$$S_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2}{2} \quad (8)$$

$$S_2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{j=1} w_{ij} + \sum_{i=1} w_{ji}\right)^2 \quad (9)$$

$$k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \quad (10)$$

The Moran I is significant and positive when the observed value of locations within a certain distance tend to be similar, negative when they tend to be dissimilar, and approximately zero when the observed values are arranged randomly and independently over space (Goodchild 1986).

3.1.2 Geary's C

Geary' s C statistic is defined as the following:

$$C = \frac{(n-1)\sum\sum w_{ij}(x_i - x_j)^2}{2\sum\sum w_{ij}\sum(x_i - \bar{x})^2} \quad (11)$$

Where x_i, x_j =the observed value at location (i, j)

$\bar{x}$ = the average of the $\{x\}$ over the n locations

w_{ij} = the spatial weight measure defined as 1 if location i and j are adjacent, or else as 0

i = contiguous to location j and 0 otherwise

The Geary statistic is always positive and asymptotically normal. The hypothesis for the Geary statistic test is that the mean of the Geary statistic is 1 if there is no spatial autocorrelation. A significant and low value (between 0 and 1) indicates a positive spatial autocorrelation while a significant and high value (greater than 1) indicates a negative spatial autocorrelation (Cliff and Ord 1981).

Spatial Pattern	*Geary's C*	*Moran's I*
Clustered Pattern	$0 < C < 1$	*I > E(I)*
Random Pattern	*C* ~= 1	*I* ~ = *E(I)*
Dispersed Pattern	$1 < C < 2$	*I > E(I)*

Table 1 Three Types of Spatial Pattern

According to Table 1(David W.S. Wong, Jay Lee 2005), the expectation stands for no spatial autocorrelation. To Geary's C, the Expectation is close to 1, that's, when C ~= 1, it illustrates the points' distribution is randomised model, when 0 < C < 1, the points' distribution is Clustered model. We consider there is probably negative spatial autocorrelation when Moran's I is smaller than the Expectation while probably positive spatial autocorrelation when Moran's I greater than the Expectation.

3.2 LISA

LISA is short for Local Indicators of Spatial Association, mainly including G Statistics, local Moran and local Geary. The G statistics (Ord and Getis 1992; Getis and Ord 1994) and LISA (Anselin 1995) provide measures for the experiments of the local spatial association. Local Moran and local Geary statistics, as suggested by Anselin (1995), are alternative local indicators. The local Moran allows for the identification of spatial agglomerative patterns similar to G statistics, while the local Geary allows for the identification of spatial patterns of similarity or dissimilarity. One advantage of the local Moran and the local Geary is that they can be associated with the global statistics (Moran I and Geary C) and can be used to estimate the contribution of individual statistics to the corresponding global statistics.

3.2.1 Local Moran

The local Moran statistic for each observation i is defined as (Anselin 1995)

$$I_i = z_i \sum_j w_{ij} z_j \quad (12)$$

where the observations z_i and z_j are the correspond deviation of x and $\bar{x}$:

$$z_i = \frac{x_i - \bar{x}}{\delta} \quad (13)$$

and w_{ij} is mostly row-standardized:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = n \quad (14)$$

The interpretation of the local Moran is similar to the G statistic (Getis and Ord 1992). A small p-value (such as $p < 0.05$) indicates that location i is associated with relatively high values of the surrounding locations. A large p-value (such as $p > 0.95$) indicates that location i is associated with relatively low values in surrounding locations.

3.2.2 Local Geary

A local Geary statistic for each observation i may be defined as follows (Anselin 1995)

$$c_i = \sum_j w_{ij}(z_i - z_j)^2 \quad (15)$$

where S 2 is the same as before, Z_i and Z_j are standardized values, and w_{ij} are the elements of a row standardized spatial weights matrix. If each observation contributes equally to the global statistic, each local Moran or local Geary should be. The

individual significance of the C_i statistic can be obtained by the same permutation approach used for Ii outlined above.

The calculation of the pseudo-significance level p-value is similar to that of local Morans. A large p-value (such as p > 0.95) indicates a small C_i in extremes, which suggests a positive spatial association (similarity) of observation i with its surrounding observations, while a small p-value (such as p <0.05) indicates a large C_i in extremes, which suggests a negative spatial association (dissimilarity) of observation i with its surrounding observations.

4. DATA AND METHODS

4.1 Data Collection

The Map data of 141 point-shaped counties and boundaries of 26 prefectures of Sichuan Region in China were attained from China Historical geographical information system (CHGIS). CHGIS is a historical GIS database on ancient China from Qin Dynasty to Qing Dynasty. It identifies the location of prefectures, and attempts to draw the boundaries of county--the sub-state administrative units, the four layers are county point file, prefecture point file, the boundaries of the regions file and the boundaries of prefecture file.

The attributes data of 141 counties and 26 prefectures mainly come from a series of 《SichuanXianZhi》 from the early days to the middle years of Qing periods, for the mass emigration began at Kangxi Period, ended at Jiaqing Period, and from some history literatures about the Toponym(Yong Lan 1995).

4.2 Data Processing and Analysis

Before the experiment, two important issues about the attribute data must be solved.

On one hand, according to the attribute data (The quantity of the Toponym) of the point layer of 141 counties, the attribute data are different from each other, though some points with these attributes were neighbours. The standardization of the column of the data should be taken because some values in the column may be 0. If the data values keep owing the original ones, it's almost impossible to calculate the true value of the global indicators, for these data appearing randomized more than normalized while the T-value is based on the normalized maximum simulation theory that need the data set to be obeying normal distribution.

On the other hand, an in-depth hotspot analysis is carried out to identify the "hotspot" and "cold spot" areas based on the logarithms (the quantity of Toponym) at "county" level. Different weight matrices are tested: inverse distance weighted, fixed distance bands of 5, 10, 15, 20 and 25 kilometers and so on. The best weight matrix is considered to be a fixed distance band of 15 kilometers. There are three reasons: first, the inverse distance weighted which fits the area observation well is too large because it may encompass all the observations in the study area while a fixed distance band of 5 kilometers is too small to encompass any observation in the study area; second, with reference to the average distance of counties which are adjacent, a fixed distance band of 15 kilometers is more reasonable; third, the experiment using the fixed distance band of 15 kilometres has a better result including relatively more significantly clustered areas.

And it's now possible to carry on the experiment following three steps: calculate the spatial autocorrelation coefficient and variance and expectation and Z value, then evaluate the statistical significance with this value, and then compare the different resulted thematic map and discover the spatial pattern of them.

5. RESULTS AND DISCUSSION

5.1 Global Moran and Geary

Global Moran and Geary (Getis-Ord general G) uses the randomization "z" statistic to evaluate the existence of clusters in the spatial arrangement of the given samples and show the level of significance with the rule that if the "z" statistic value is greater than the key value 1.96 then we consider the significance level of the given samples is 5%, and then if the "z" statistic value being even greater than another key value 2.576 will lead to a higher significance level of 1%.
In the experiment, we normalized the quantity of the Toponym in each of 141 counties and each row of the spatial weight matrix and chosen 15km as the fixed band distance, and calculated Moran's I and Geary's C values in the table below:

	I or *C*	*E(I)* or *1*	*Variance*	*Z Value*
Moran's I	0.0532	-0.0071	0.0002	4.8743
Geary's C	0.5799	1	0.0110	2.7629

Table 2 The Values of Global Moran's I and Geary's C

From Table 2 we see that the "z" statistic value of Moran's I is 4.8743, which is bigger than 2.576, that's the significance level is 1%, showing us that the distribution pattern of the Toponym of Sichuan region after the mass emigration is a high cluster and has high spatial autocorrelation. The Geary's C value in Table 2 also shows the existence of strong spatial autocorrelation in the research area with C value 0.5799 and "z" value 2.7629.

5.2 Hot spot and cold spot analysis（G_i）

Being a composite index, Global Moran and Geary is the measure of the overall clustering of the data, used to evaluate the overall spatial association of the total research area. But it is reasonable for us to consider that the spatial autocorrelation level of different census area is not exactly the same. For this reason, we use a local indicator called Gi (Getis-Ord Gi*) to detect and evaluate the spatial autocorrelation of local census area, high Gi means the census area is a cluster of high ratio while low Gi means the census area is a cluster of low ratio.

During the experiment, we have standardized the quantity of the Toponym in each of 141 counties, standardized each row of the spatial weight matrix, chosen 15km as the fixed band distance, and calculated Local G_i values which are visualized as the point layer, and classified the points in seven kinds which are marked different colours according to Z value range in legend "Couty_Point" in Figure 1, meanwhile, we have standardized the quantity of the Toponym in each of 26 prefectures, and built their spatial weight matrix calculated by

the inverse weighted distance, standardized each row of it, and calculated G_i data which are visualized as the area layer, and classified the polygons in seven kinds which are marked different colours according to Z value range in legend "Prefecture_area" in Figure 1 below:

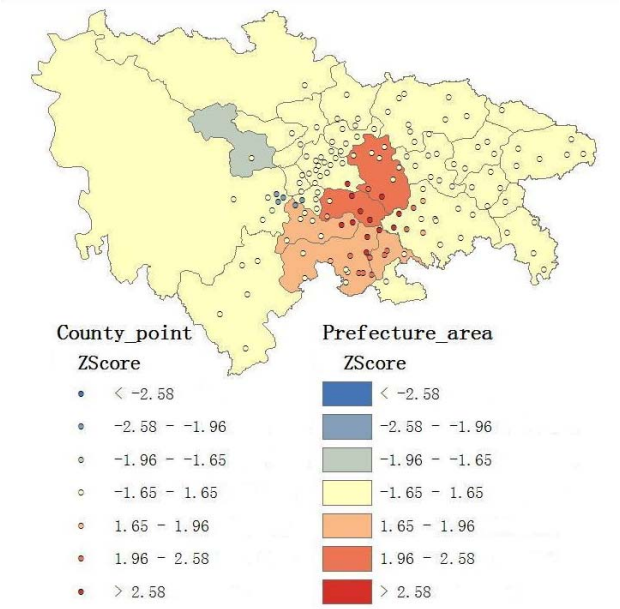

Figure1. Thematic Map Using Local G-Statistic Method

We use a local indicator called G_i (Getis-Ord G_i^*) to detect and evaluate the spatial autocorrelation of local census area, high G_i means the census area is a cluster of high ratio while low G_i means the census area is a cluster of low ratio.

Global Moran and Geary (Getis-Ord general G) uses the randomization "z" statistic to evaluate the existence of clusters in the spatial arrangement of the given samples and reveals the level of significance with the rule that if the "z" statistic value is greater than the key value 1.96 then we consider the significance level of the given samples is 5%, and then if the "z" statistic value being even greater than another key value 2.576 will lead to a higher significance level of 1%.

Figure 1 shows us that the similar points with high ratio in red clustered in the red area, while the points with sub-high ratio in orange clustered in the orange area. That's, the hot spot area of the point layer matches well with the hot spot area of the prefecture polygon layer. And it further revealed that the majority of counties in the three red prefectures lies in the south of Sichuan region had a highly developed emigration culture after the ancient mass emigration.

5.3 Local Moran Analysis

Global spatial autocorrelation analysis yields only one statistic to summarize the total research area. In other words, global analysis assumes spatial homogeneity. As we all know that spatial heterogeneity is one type of spatial association which is a basic feature in geographic researches so that assumption does not hold, then having only one statistic does not make sense as the statistic should differ over space. Furthermore, we can still find clusters at a local level using local spatial autocorrelation if there is no global autocorrelation or no clustering.

During the experiment, we have standardized the quantity of the Toponym in each of 141 counties, standardized each row of the spatial weight matrix, chosen 15 km as the fixed band distance, and calculated Local Moran's I values which are visualized as the point layer, and classified the points in seven kinds which are marked different colours according to Z value range in legend "Couty_Point" in Figure 1, meanwhile, we have standardized the quantity of the Toponym in each of 26 prefectures, and built the their spatial weight matrix calculated by the inverse weighted distance, standardized each row of it, calculated Local Moran's I values which are visualized as the area layer, and classify the polygons in seven kinds which are marked different colours according to Z value range in legend "Prefecture_region" in Figure 1 below:

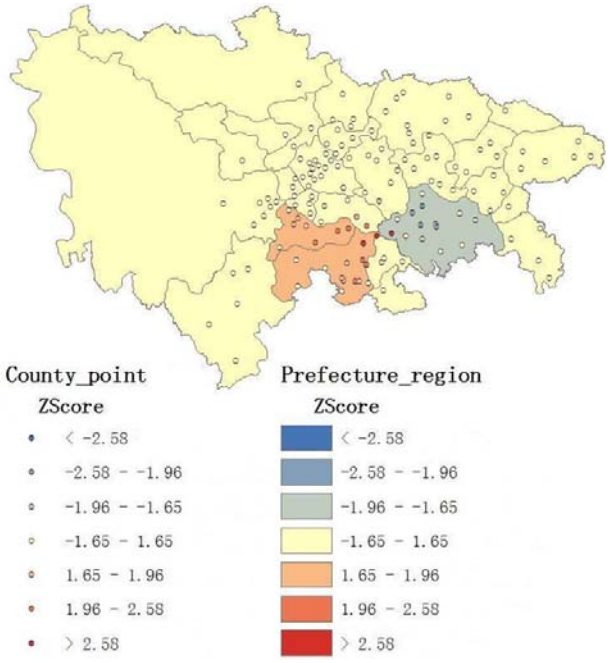

Figure2. Thematic Map Using Local Moran Mean

Local Moran Indicator just evaluates the similarities and dis-similarities of clusters and it cannot tell us whether the degree of the clusters is great or small. In Figure 2, red points denote there are points with similar ratios clustering together and the red region in the south also matches well with the hot spot area in Figure. On the contrary, blue points in the blue areas in the southeast reflect the majority of the counties of the prefecture have dissimilar emigration culture development patterns from one another.

6. CONCLUSION

First, the Global Moran and Geary analysis proved strong spatial autocorrelation in the spatial distribution of the Toponym of Sichuan region which showed that one point may be surrounded by other points with similar attributes to itself in other words, the closer one point to another the more similar the two points are. It could be helpful in the Cultural Geography research work taking spatial association into consideration; Second, the hot spot analysis shows spatial association of the Toponym with hot spot area in red and cold spot area in blue, which further indicates the existence of sub-areas that

developed differently over space in the cultural domain. In the emigration cultural research area, We tentatively interpret the hot spot as the most developed emigration culture areas and the cold spot as the most undeveloped; Third, Local Moran analysis illustrates that the most developed culture areas in the cultural research region tend to be a cluster which indicates that the spatial pattern of cultural resources distribution does exist with similar features clustering together.

The application of spatial analysis methods based on ESDA in analyzing the spatial distribution patterns of cultural resources has shown us that it is possible to use spatial statistic methods in cultural domain and proved that geographic location factors and spatial associations could be used in Cultural Geography research. These two thematic maps before-mentioned are illustrations of spatial autocorrelation of the distribution of the ancient Toponyms in Sichuan region after the mass emigration. Spatial statistic methods appeared therefore as a powerful tool to reveal the characteristics of cultural regions in sub-administrative unit (i.e. prefecture) is in relation to those of its geographical environment and yield scientific explanations for spatial distribution patterns.

REFERENCE

Zhengxiang Cheng, 1983. *China Cultural eography • Taiwan's Toponym* — Neolithic cultures.Sanlian Bookstore.

Shangji Situ, 1992. The Historical geographical Research of GuangDong's Toponym. *Comments on Chinese Historical Geography*.

Yong Lan, 1995. The Research on Geographical features of the Distribution of Local Residents and Emigration in Qing period in Sichuan Region. *Comments on Chinese Historical Geography*.

Fahui Wang, 2009. GIS based quantitative methods and their applications. *The Commercial Press*, pp. 51~61, 208~210.

Anselin, L., 1995. Local Indicators of Spatial Association - LISA, *Geographical Analysis* 27, pp. 93-115.

Bao, Shuming and Mark S. Henry. 1996. Heterogeneity issues in local measurements of spatial association. *Geographical Systems*, Vol. Ⅲ, pp: 1-13.

Cliff A. and Ord, J.K. 1973. *Spatial Autocorrelation*. Pion, London.

Cliff, A. D. and Ord, J. K., 1981. *Spatial Processes: Models and Applications*. Pion, London.

Cressie, Noel A, 1993. *In Statistics for Spatial Data*, John Wiley & Sons, Inc. pp. 79-122.

David W.S. Wong, Jay Lee, 2005. *Statistical Alalysis of Geographic Information with Arcview GIS and ArcGIS*, John Wiley & Sons, Inc. pp. 302-367.

Cressie, Noel A. and Hawkins, D.M., 1980, Robust estimation of the variogram, *I.Journal of the International Association for Mathematical Geology*, 12, pp. 115-125.

Getis, Arthur and Ord, J. Keith. 1996. Local Spatial Statistics: An Overview. In Spatial Analysis: Modeling in a GIS Environment, *Geoinformation International*. P. Longley and M. Batty (eds.), Cambridge, UK.

Getis, Arthur and Ord, J. Keith. 1995. The Use of a Local statistic to Study the Diffusion of AIDS from San Franciso, *Regional Science AssociationInternational in Cincinnati*.

Getis, Arthur and Ord, J. Keith. 1992. The Analysis of Spatial Association By the Use of Distance Statistics. *Geographical Analysis*, 24, pp. 189-206.

Goodchild, M. F., Haining, R. P. and Wise, S. 1992. Integrating GIS and spatial data analysis: problemsand possibilities. *In International Journal of Geographical Information Systems* 6(5), pp. 407-423.