Spatial Clusters of Diseases: Remodeling the Concept

Ge Lin* and Donglin Zeng†

*Department of Geography, University of Victoria
Victoria BC CANADA, V8W 3P5
†Department of Statistics, University of Michigan
Ann Arbor, MI 48109, USA

Abstract
In this article, we propose an alternative way of testing spatial clustering for common diseases. In order to detect a hot spot, we treat a global cluster statistic from a localized perspective, and define an area with positively correlated neighboring regions as a cluster. The proposed test uses the maximum likelihood method to detect the existence of a cluster, and it does not require the calculation of the mean and variance as most spatial statistic tests do. Using the spatial chi-square test of Rogerson (*R*) as a benchmark, our subsequent simulations and case study show that when the existence or nonexistence of spatial clusters are apparent, our test result is consistent with R. However, when a low value region surrounded by high value neighbors is considered, the result from R finds a cluster, while our result finds no cluster.

I. INTRODUCTION

In the past few years, several new methods (Getis and ord, 1992; Anselin, 1995) have been proposed to measure spatial associations for cluster analyses. These tests are based on the test of spatial autocorrelation (e.g., Moran *I*), which assumes that either attribute values (e.g., disease prevalence) is in equal probability among all the geographic units or from a single parent distribution. However, many have noted that population sizes often vary substantially between rural and urban units, and when the traditional permutation test of equal probability applies to this situation, substantially large variation often occurs in sparely populated areas. For the global test, Oden (1995) applied regional population weights to Moran *I* statistic to adjust for the regional distributions of diseases. For the local test, Bao and Henry (1996) provided a generalized form of local spatial statistic (*GLISA*) that also accounts for population distribution in the study area. These developments are examples where usual statistical methods are modified to take into account the spatial autocorrelation in the existing data structure (Dutilleual, 1993). Consequently, if one wants to determine the existence of a spatial cluster, one has to make the judgement at a local scale if the significant positive local association could be considered a cluster.

Parallel to these developments, several attempts are also made to work on the underlying distribution of spatial dependency of disease data. Using the χ^2 approximation with the degree of freedom being adjusted by a Gamma function, Tango (1995) proposed a general test, C_g, to determine whether spatially distributed disease rates are independent or clustered based on a spatial weight (*W*) matrix. Similar to Moran *I*, an expected rate is derived for each region within a study area and this rate is compared with the observed rate within each region. Similar to Oden, Tango's C_g also adjusts for population size in each region. A special case of Tango C_g is the Rogerson *R* (Rogerson, 1998; 1999), a spatial version of the Chi-square goodness-of-fit statistic.

Given a population size (ξ_t) and disease prevalence (N_i) at region i for a study area with m regions, the random variable $\boldsymbol{r}$ is the $m \times 1$ vector of $r_i = N_i/N$, where $N=N_1+N_2+\ldots+N_m$, and the nonrandom variable $\boldsymbol{p}$ can be expressed by the $m \times 1$ vector of $p_i=\xi_i/\xi$, where $\xi=\xi_1+\xi_2+\ldots+\xi_m$. The spatial chi-square goodness-of-fit statistics is defined as:

$$R= \sum_j \sum_j w_{ij} (r_i-p_i)(r_j- p_j) \quad (1)$$

Where w_{ij} are elements of the weight matrix (W) defined by

$$w_{ij} = a_{ij} / \sqrt{p_i p_j} \ . \quad (2)$$

a_{ij} is a measure of geographic closeness of region i to region j. Substitute w_{ij} with $a_{ij} / \sqrt{p_i p_j}$ in equation (1) we have

$$R= \sum_j (r_i-p_i)^2/p_i + \sum \sum_{i \neq j} w_{ij} (r_i-p_i)(r_j- p_j) / \sqrt{p_i p_j} \quad (3)$$

Rogerson further provides the expected value and variance of R, and the test of significance using the chi-square approximation with the degree of freedom being adjusted by a Gamma function. It is clear from equation (3) that the first term is the usual chi-square statistic and the second term is the aspatial chi-square statistic, and R is the sum of the two terms. However, equation 3 also reveals some problems. First, either

1082-4006/99/0502-175$5.00

'hot' or 'cool' spots would make the test significant just as the usual chi-square test does. Second, cool and hot spots could coexist in a disease pattern, and these tests cannot differentiate a cool spot from a hot spot. As a result of second problem, these tests may not be sensitive to local variation within a cluster, thus a hot spot could contain a cool point or verse versa. In other words, these tests generally only provide statistic significance for the existence of spatial associations similar to *LISA* and *G* statistics at the local level.

In order to determine if a potential spatial association around a neighborhood can be treated as a cluster, we propose an alternative global test. Similar to Tango's C_g or Rogerson's R, we try to identify disease clusters or any other spatial clusters associated with rates. However, a spatial cluster in our case is primarily based on the positive correlation among local values (rates), meaning a hot spot would not contain a cool point. In the following section we provide analytical and statistical procedures for the measure, and perform the test with simulated data. Results from our test are compared with those from Tango's and Rogerson's tests. In section 3, we provide a case study about spatial distributions of elderly disability in Alabama and Mississippi. Finally, we discuss the alternative statistic with some concluding remarks.

II. SPATIAL CLUSTERS AS CORRELATED NEIGHBORS

Regardless of the nature of problems (e.g., continuous or discrete), a common feature of most spatial statistics, according to Cliff and Ord (1981), is that they are similar in the approach to the t-test (or ANOVA). Spatially connected pairs are measured for their similarity by referring to the sample mean, and this similarity is then contrasted with a measure of spatial similarity (e.g., spatial weight matrix). This type of test is a natural extension of the case where there is no spatial autocorrelation. In essence, it is a test for spatial associations rather than for a cluster. As noted by Anseling (1995), even when there is no spatial autocorrelation, local clusters can still exist. To test the existence of a cluster following this thread, we do not have to compare local values with the expected values. We can instead focus on similarity in values around a neighborhood. Hence, contrary to common spatial autocorrelation tests of no difference in mean, we can construct a spatial structure with a built-in cluster component for the existing data, and then test the significance of the cluster component against the null hypothesis of no cluster.

This analytical framework is similar to the parameterized simultaneous autoregressive (SAR) tests (Haining, 1990) against an existing pattern. It assesses the spatial dependence by testing the strength of correlation among neighbors for each location. If values in neighboring regions for a given location are positively correlated, then regions around that location are clustered. Such a test is very close to the interpretation of G statistics (Getis and Ord 1992) when there is a positive spatial association. This framework is particularly useful for relating environmental factors to a disease pattern. For instance, in searching for environmental factors causing certain diseases, environmental health specialists and epidemiologists might be interested in a cluster where every region in a hypothetically clustered area has an excessive rate. This could imply that either the whole cluster area is contaminated or embedded with some environmental deficiency. However, if some regions were excessive, while others were not, it would be difficult to convince environmental health specialists that an environmental factor causes the excessive prevalence of the disease in that particular area. Thus, unlike general cases, where spatial associations can be positive or negative, the spatial cluster is measured by the correlation of neighborhood values. Formally, a spatial cluster is a location/region with positively correlated neighboring regions. According to this definition, a region with a low value or rate surrounded by regions with high values cannot be viewed as having a cluster, and a possible negative spatial association is not defined.

Following Rogerson's notation, there are m regions in a study area. Let ξ_i and N_i denote, respectively, the population size and the number of disabled persons at region i. If no spatial association exists, N_i are independent Poisson variables, whose distributions can simply be approximated by a normal distribution (Johnson, 1982). In a more general case, assuming p is the probability of each person being disabled in the study area, we have a set of random Poisson variables N_i, each with a mean of $p\xi_i$. These Poisson variables asymptotically follow a multinormal distribution, which is an extension of the normal distribution in a multi-dimension space for correlated dependent variables, i.e.,

$$\frac{N_i - p\xi_i}{\sqrt{\xi_i}} \Rightarrow N(0,\sigma^2)$$

where $\sigma^2=p$. This formulation, similar to Oden's adjusted Moran *I*, implicitly accounts for the population size in each region. Based on the assumption of large population ξ_i, it is reasonable to assume that

$$B_i = \left(\frac{N_i - p\xi_i}{\sqrt{\xi_i}}\right), \quad i = 1,\ldots,m \quad and \quad B = \begin{pmatrix} B_1 \\ \ldots \\ B_m \end{pmatrix} \qquad (4)$$

has a multinormal distribution $N_m(0, \sigma^2(I_m+\varepsilon W_m))$ with the mean of zero and covariance matrix of $\sigma^2(I_m+\varepsilon W_m)$. Here, the covariance structure has a similar model structure as Rogerson's R. If there is no spatial cluster, the ε term would be zero, leaving the usual σ^2 term. W_m is a spatial weight matrix indexed by i and j, and the subscript m indicates the number of regions or the number of rows in the matrix. When calculating the covariance matrix, how to construct W_m is entirely at our disposal. A W_m can be a usual zero-one weight matrix, or a continuum defined by some distance measure between region i and region j. In both Tango and Rogerson, an exponential distance function, that is, $\exp(-d(ij)/\tau)$ was used. If we let $1/\tau$ equal β, it becomes the usual "distance decay" parameter with larger values being associated with shorter distance influence and smaller values being associated with longer distance influence around each location. This parameter can be tuned to see the impact of "distance frictions" in various spatial scenarios.

Under the null hypothesis (H_0) of independent spatial distribution, ε should be statistically close to 0. One way to test is to construct a likelihood ratio test comparing the observed pattern with the pattern under the null hypothesis (Casella, 1990). If the ε makes a difference between the observed and the one under the H_0, then the likelihood ratio statistic should be big enough so that the chance for the observed pattern to match the H_0 pattern is very small (e.g., $p<0.05$). More precisely, testing the existence of clusters is equivalent to testing: $H_0 : \varepsilon=0$ vs. $H_1 : \varepsilon>0$.

Although there might be a closed distribution form for the observed pattern, we do not have to derive it when using the log-likelihood ratio test. If the largest log-likelihood ratio statistic is small, then the observed pattern is similar to the pattern of no spatial cluster. A measure of spatial cluster, therefore, can be obtained as the difference between $2\max_{\varepsilon>0}$Log-likelihood and $2\max_{\varepsilon=0}$ log-likelihood. After working out some algebra (APPENDIX I), we derive our testing statistics as:

$$T = -\inf_{\varepsilon>0}[m \ln B^T (I_m + \varepsilon W_m)^{-1} B + \ln | I_m + \varepsilon W_m |] + m \ln(B^T B) \quad (5)$$

Where, $\inf_{\varepsilon>0}$ refers to the minimization over all positiveε, hence its negative is a maximization process. $\hat{p}$ is the maximum likelihood estimator of p given by $\sum N_i / \sum \xi_i$. When the likelihood function achieves the maximum, we have $\hat{p}$, $\hat{\sigma}$ and the corresponding ε for our test T, which asymptotically follows Chi-square distribution with one degree of freedom. A large T suggests that the observed distribution differs from the independent distribution, and the H_0 is likely to be rejected. A small T, on the other hand, is likely to accept the H_0.

To see how the statistic performs, we generated a 10 by 10 grid (in 10 miles each unit) and obtained Ts with three simulated patterns. First, a randomly generated population exposure along with the number of disabled persons was randomly assigned to each region. The population ranges from 300 to 500 in each region, and the number of disabled persons ranges from 20 to 60. Euclidean distances between each grid are used in the exponential distance function identical to Tango, where $\tau=1$ or $\beta=1$. The P-value for our test is close to 1, for Tango's C_g is 0.15 and for Rogerson's R is 0.12, suggesting that all the tests yield statistically similar results when there is no spatial cluster. Next, we raised the number of people with a disability randomly by 25 to 30 around the central 3 by 3 grids. In this case, all three tests are significant with p-values close to 0 ($P<0.05$), suggesting that all of them are capable of detecting a simulated cluster for a centrally located "hot spot." Finally, we sank the central location within the 3 by 3 grids to a low value while keeping all other values from the previous scenario. The result from our test accepts the H_0 of no clusters, but the results from C_g and R reject the H_0. It suggests that C_g and R are not sensitive to any continued plateau even with a basin at the center, whereas our test treats a low value surrounded by high value neighbors as no cluster.

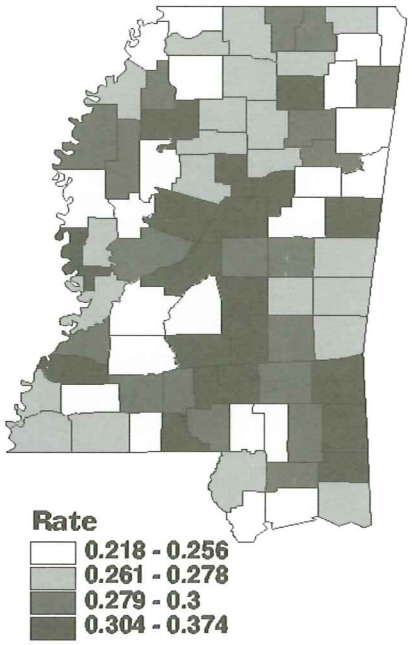
III. CASE STUDY

Spatial patterns of elderly disability in Alabama and Mississippi. At the national level, the spatial disparity of elderly disabilities has a very distinct pattern: a high concentration in the Deep South and low concentration in the Midwest (Lin, 2000). In this case study, we select Alabama and Mississippi, the two states with the highest elderly disability rates. In order to establish spatial relationship between elderly disability with other spatial processes, it is necessary to determine if there is any disability cluster in these states. If there is no spatial cluster, then socio-environmental factors in the south at a large geographic scale may contribute to the southern concentration of elderly disability. If, on the other hand, there are some spatial clusters, then some localized socio-environmental processes may contribute to the state level concentration, and we will want to correlate these processes with the clusters.

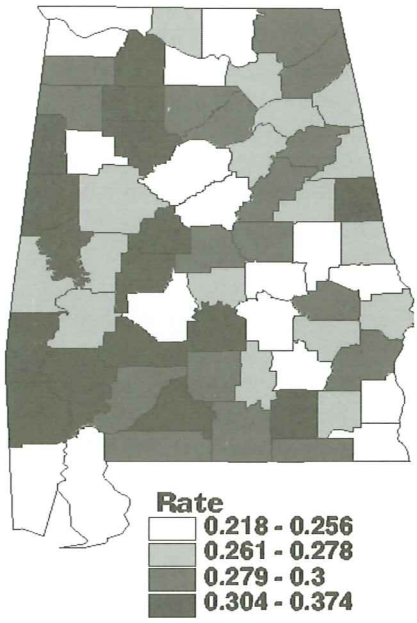
Elderly disability rates at the county level are shown in Figure 1a, 1b, and 1c (Figure 1). These summary data are based on the long-form of the 1990 census[1].

The elderly is defined as those who are 65 and over, and a person is considered having a disability if she or he has a mobility or self-care limitation. When county-level disability rates for Alabama (1a) and Mississippi (1b) are viewed separately, it is hard to find any high or low concentrations of disability. If the rates for the two states are combined (1c) in a single map, there are some indications of spatial clusters or concentrations leaning toward Mississippi.

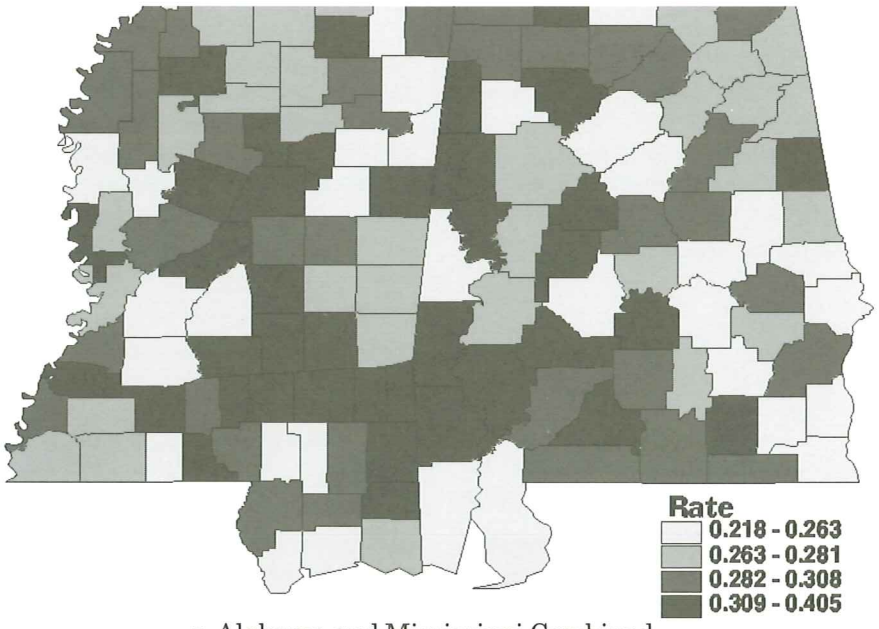
We performed three analyses for the three spatial patterns corresponding to Figure 1a, 1b, and 1c, that is, examine disability distributions separately for Alabama and Mississippi, and their combined distribution. Results from both C_g and R (Table 1, first and second columns) indicate the existence of spatial clusters for all the patterns. For our test using the identical exponential function, there is no spatial cluster for Alabama or Mississippi if the two states are investigated separately; there is an indication of the cluster if the two states are jointly evaluated. As the ex-



𝔹. Mississippi Alone



a.Alabama Alone



c. Alabama and Mississippi Combined

Figure 1. Elderly disability rates in Alabama and Mississippi

[1]Statistics are generated using S-Plus software. Any data and program codes used in this study are available upon request.

Table 1. P-Values for testing spatial clusters for Mississippi and Alabama

	Tango Cg	Rogerson R	T(exp)[a]	T(power, $(1+\alpha d_{ij})^{-\beta}$)	
	τ=0.01	τ=0.01	β=100	β=1, α=0.2	β=1,α=1
MS	0	0	0.13	.18	0.22
AL	0	0	0.20	.25	0.24
MS & AL	0	0	0.02	.003	0.03

a: The distance function in T(exp) is identical to the one used for calculating R or Cg.

ponential function is more sensitive to changes in distance measurements than the power function, we also used several power functions to reevaluate these patterns (Table 1, last two columns). The formula for the power function is $(1+\alpha d_{ij})^{-\beta}$, where α is a scale parameter for adjusting distance measurements, and β is a tuning parameter for distance decay effects. This power function equals 1 for the diagonal elements of the W, which is identical to the value in the exponential function used by Rogerson. We experimented with this power function with several levels of α and β. In general, if β is tuned "appropriately," the results are fairly consistent with different αs, rejecting H_0 if the two states are combined, and accepting H_0 if the two states are assessed separately. However, when β >2, which corresponds to considering a smaller neighboring region in the model, our results with different αs (e.g., α=0.2; 1; 1.5) tend to accept H_0. This is understandable, since the most influential areas under the examination are so few that they can hardly constitute a cluster. We also experiment with different C_gs or Rs with a power function. The results are similar, rejecting H_0 for all three cases. For example, when $(1+\alpha d_{ij})^{-\beta}$ with α=0.2 and β=1 is used, C_g has the 163.42 likelihood ratio chi-square with 2.55 degrees of freedom, as opposed to 263.77 likelihood ratio chi-square with 3.88 degrees of freedom given the exponential function.

There are several reasons why our results differ from C_gs or Rs in this case study. Most importantly, our test by definition is less sensitive to some spatial associations, as it excludes some cases (e.g., hot spots with some cool points), where a spatial association might exist, but the rates might not be positively correlated or excessive for all areas around a neighborhood. Secondly, the likelihood ratio test only searches the largest likelihood ratio statistic (with one degree of freedom), whereas the R is based on the cumulative deviations adjusting for the degrees of freedom. Thirdly, it seems that very small variances resulting from Tango's or Rogerson's formulations make a huge difference (see Tiefelsdorf and Boots, 1997 for a discussion of the variance in a similar context). Finally, our normal approximation may not always appropriate when the sample size in a particular county is not sufficiently large.

IV. SUMMARY

In summary, we have introduced a test for spatial clusters of cool or hot spots, and compared it with Tango's Cg and Rogerson's R. One significant property of this method is that it primarily models the neighboring covariance instead of neighboring means, thus requiring no calculation of expected value and variance. Compared with R, our test is more sensitive to local variation in values (rates), and it tends to treat a low value region surrounded by high-value neighbors as no cluster, while other tests may find an existence of a cluster. Computationally, our test is more intensive and time consuming than R, because it needs to search the maximum of the log-likelihood.

Even though C_g, R, and T depend on distance scales, the underlying notion is not unreasonable: "the exposure relates inversely to some geographical distance from the focus" (Tango, 1995, p. 2324). However, there is some arbitrariness of tuning the β or τ parameter, as β cannot be calibrated in the absence of other spatial information. Perhaps several βs should be used to reflect different spatial processes.

Finally, despite the global nature of the test, it is possible to approximately decompose the statistic (T) into the sum of local contributions to the global statistic by partitioning the W matrix into m local weight matrices. A greater contribution of a location i indicates the greater possibility of a cluster around that location. And it is possible to use this local indicator to detect clusters around a specific region. In addition, we can also test the statistic power of T by providing a set of εs. Our next step is to explore ways of measuring local clusters and testing the statistic power of the T.

APPENDIX I. The Derivation of Likelihood RatioTest T

Given the density function of the multinormal distribution

$$f(X) = \left|2\pi\sigma^2(I_m + \varepsilon W_m)\right|^{-\frac{1}{2}} \exp\{-\frac{1}{2} B^T (I_m + \varepsilon W_m)^{-1} B\}$$

we can construct a likelihood ratio test with $(\mathrm{Max}_{Ho} f(X))/(\mathrm{Max}_{H1} f(X))$ or its equivalent test of the 2loglikelihood=$2l(\varepsilon,\sigma,p)$

$$= -2m\ln\{\sqrt{(2\pi)}\sigma\} - \ln|I_m + \varepsilon W_m| - \frac{B^T(I_m + \varepsilon W_m)^{-1}B}{\sigma^2}$$

to maximize$2l(\varepsilon,\sigma,p)$, $\hat{p}$ is the least square estimator of regressing $N_i/\mathrm{sqrt}(\xi_i)$ on $\xi_i/\mathrm{sqrt}(\xi_i)$. Therefore, $\hat{p} = \sum N_i / \sum \xi_i$. Differentiating $2l(\varepsilon,\sigma,p)$ over σ, we have

$$0 = \frac{-2m}{\hat{\sigma}} + \frac{2}{\hat{\sigma}^3} B^T(I_m + \varepsilon W_m)^{-1}B$$

$$\Rightarrow \quad \hat{\sigma} = \sqrt{\frac{B^T(I_m + \varepsilon W_m)^{-1}B}{m}}$$

Substituting $\hat{p}$ and $\hat{\sigma}$ into the log-likelihood function, we have

$$2l(\varepsilon, \hat{\sigma}, \hat{p}) = -m\ln(2\pi) - m\ln(B^T(I_m + \varepsilon W_m)^{-1}B) + m\ln(m) - \ln|I_m + \varepsilon W_m| - m$$

Our statistic

$$T = \max_{\varepsilon>0} 2l(\varepsilon,\hat{\sigma},\hat{p}) - 2l(0,\hat{\sigma},\hat{p})$$
$$= \max_{\varepsilon>0}(-m\ln B^T(I_m + \varepsilon W_m)^{-1}B - \ln|I_m + \varepsilon W_m|) + m\ln(B^T B)$$
$$= -\inf_{\varepsilon>0}[m\ln B^T(I_m + \varepsilon W_m)^{-1}B + \ln|I_m + \varepsilon W_m|] + m\ln(B^T B).$$

ACKNOWLEDGEMENT

We are grateful to Peter Rogerson, who provided the data and GAUSS codes for this study. These data and codes enabled us to replicate Rogerson's work and to formulate our test. We would also like to thank Julie Zhou and Shuming Bao for helpful comments on an early draft.

REFERENCES

[1] Anseling, L., 1995, Local indicators of spatial association—LISA, *Geographic Analysis,* 27:93-115.

[2] Bao, S, and Henry, M. S., 1996, Heterogeneity issues in local measurements of spatial association, *Geographic Systems,* 3:1-13.

[3] Casella, G., 1990, *Statistical inference*, Pacific Grove, CA.

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

[5] Dutilleual, P., 1993, Modifying the t test for assessing the correlation between two spatial processes, *Biometrics* 49: 305-314.

[6] Getis, A., and J. Ord, 1992, The Analysis of spatial association by use of distance statistics, *Geographical Analysis,* 24:189-206.

[7] Haining, Robert, 1990, *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press, Cambridge, UK.

[8] Johnson, A. Richard, 1982, *Applied Multivariate Statistical Analysis,* Englewood Cliffs, N. J.: Prentice-Hall.

[9] Lin, G, 2000, The Geographic assessment of elderly disability in the U.S, *Social Science & Medicine*, 50(7-8): 1015-1024

[10] Oden, N., 1995, Adjusting Moran's I for population density, *Statistics in Medicine,* 14:17-26.

[11] Rogerson, P. A., 1998, A spatial version of the chi-square goodness-of-fit test and its application to tests for spatial clustering, In *Econometric Advances in Spatial Modeling and Methodology: Essays in Honor of Jean paelinck,* Eds C. Amrhein, D. Griffith, and J.M. Huriot. Dordrecht: Kluwer,71-84.

[12] Rogerson, P. A., 1999, The detection of clusters using a spatial version of the chi-square goodness-of-fit statistics, *Geographical Analysis* 31:130-147.

[13] Tango, T., 1995, A class of tests for detecting 'general' and 'focused' clustering of rare diseases, *Statistics in Medicine* 14: 2323-2334.

[14] Tiefelsdorf, M. and Boots, Barry, 1997, A note on the extremities of local moran's I's and their impact on global moran's I, *Geographical Analysis,* 29:249-257.