Summer 2021

A Comparison of Spatial Clustering Assessment Methods

Nadeesha Dilhani Vidanapathirana

A COMPARISON OF SPATIAL CLUSTERING ASSESSMENT METHODS

By

Nadeesha Dilhani Vidanapathirana

Bachelor of Science
University of Colombo, 2016

Master of Applied Statistics
Louisiana State University, 2019

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatistics

Arnold School of Public Health

University of South Carolina

2021

Accepted by:

Stella Self, Director of Thesis

Yuan Wang, Reader

Alexander McLain, Reader

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Stella Self for her assistance at every stage of my research project. Without her support, it would be impossible for me to complete this study in such a short time. I am deeply grateful to Dr. Yuan Wang for her consolation and encouragement in the past few months. I would also like to thank Dr. Alexander McLain for his insightful comments and suggestions. My appreciation also goes out to my parents and my husband for their unwavering support and belief in me.

ABSTRACT

Spatial clustering detection methods are widely used in many fields of research including sociology, epidemiology, ecology, and criminology. The objective of this study is to assess the performance of four spatial clustering detection methods: the average nearest neighbor ratio, Ripley's K function, local Moran's I and Getis-Ord Gi* statistics. We conduct a simulation study to evaluate the performance of each method for areal data under different types of spatial dependence and three different areal structures; a 20x20 regular grid, United States counties in six states and Canadian forward sortation areas (FSAs) in three provinces. The results shows that the empirical type I error rates are inflated for ANN and Ripley's K. For local Moran's I and Getis- Ord Gi* statistics empirical type I error rates are less than or equal to 0.05 for most of the units in all three areal structures and classification accuracy is closer to 1. We find that the performance of ANN and Ripley's K are not reliable when applied to areal data unlike local Moran's I and Getis-Ord Gi*.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

ABSTRACT iii

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER 1: Introduction 1

CHAPTER 2: Methods 4

2.1 Average Nearest Neighbors 5

2.2 Ripley's K function 7

2.3 Local Moran's I 10

2.4 Getis-Ord Gi* 12

CHAPTER 3: Simulations 14

CHAPTER 4: Results 24

4.1 ANN and Ripley's K 24

4.2 Local moran's I and Getis-Ord Gi* 26

CHAPTER 5: DISCUSSION 38

CHAPTER 6: CONCLUSION 41

References 42

LIST OF TABLES

Table 4.1 Type I error and power of ANN. $n_a = 267$ for Canadian FSAs, $n_a = 400$ for regular grid, $n_a = 549$ for USA counties. 24

Table 4.2 Type I error and power of ANN with different windows for each simulation of data. $n_a = 267$ for Canadian FSAs, $n_a = 400$ for regular grid, $n_a = 549$ for USA counties. 25

Table 4.3 Type I error and power of Ripley's K for 5 different radii (20x20 regular grid). 25

Table 4.4 Type I error and power of Ripley's K for 5 different radii (USA counties)..... 26

Table 4.5 Type I error and power of Ripley's K for 5 different radii (CA FSAs). 26

Table 4.6 Mean Type I error for local Moran's I (LM) and Getis-Ord Gi* (GG*). 35

Table 4.7 Mean classification accuracy for local Moran's I. 36

Table 4.8 Mean classification accuracy for Getis-Ord Gi*. 37

LIST OF FIGURES

Figure 2.1 Example of weighted edge correction, dij is the distance between the ith and jth points, e is the distance between point i and the nearest boundary, the value of α gives the inverse cosine function in Eq. 4 (Haase, 1995). 9

Figure 2.2 Examples of spatial patterns of clustering (left), CSR (middle), and dispersion (right). 9

Figure 2.3 Cluster/outlier types. Red indicates high values and blue indicates low values. 11

Figure 3.1 Observed units for 20x20 regular grid (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes. 16

Figure 3.2 Observed units for USA counties (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes. 17

Figure 3.3 Observed units for CA FSAs (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes. 18

Figure 3.4 Observed values generated for 20x20 regular grid under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers. 19

Figure 3.5 Observed values generated for USA counties under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers. 20

Figure 3.6 Observed values generated for CA FSAs under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers. 21

Figure 4.1 Type I error of Local Moran's I (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high

clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers. 28

Figure 4.2 Type I error of Getis-Ord Gi* (20x20 regular grid) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers. 29

Figure 4.3 Type I error of Local Moran's I (USA counties) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers. 29

Figure 4.4 Type I error of Getis-Ord Gi* (USA counites) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers. 30

Figure 4.5 Type I error of Local Moran's I (CA FSAs) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers. 31

Figure 4.6 Type I error of Getis-Ord Gi* (CA FSAs) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers..... 31

Figure 4.7 Classification accuracy of local Moran's I (20x20 regular grid) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers. 32

Figure 4.8 Classification accuracy of local Moran's I (USA counties) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low

outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers. 32

Figure 4.9 Classification accuracy of local Moran's I (CA FSAs) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers. 33

Figure 4.10 Classification accuracy of Getis-Ord Gi* (20x20 regular grid) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers..... 34

Figure 4.11 Classification accuracy of Getis-Ord Gi* (USA counites) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers..... 34

Figure 4.12 Classification accuracy of Getis-Ord Gi* (CA FSAs) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers..... 35

CHAPTER 1: INTRODUCTION

Researchers have been using spatial clustering analysis for many years to analyze data for spatial patterns. One of the earliest examples of spatial clustering analysis in public health occurred in 1894 when Dr. John Snow mapped the location of cholera cases to identify the source of an outbreak in London and was able to identify a particular water pipe as the source (Moore & Carpenter, 1999). Over the past few decades, the popularization of geographical information systems (GIS) software has fueled the development of many new methods of spatial analysis.

Spatial clustering analysis techniques are used in various disciplines including sociology, epidemiology, meteorology, and ecology. Often when we apply spatial clustering analysis, we seek to answer questions such as, 'are the locations of observed data spatially clustered, or are they random?' or 'do similar observations tend to be closer together?'.

Spatial clusters can be identified from point process data or areal data and several statistical methods have been developed to analyze spatial patterns in these data. Here, we are going to consider four such methods; the average nearest neighbor (ANN) ratio, Ripley's K function, local Moran's I statistic, and Getis-Ord Gi* statistic. In 1954, Clark and Evans (Clark & Evans, 1954) used the distance from an individual to its nearest neighbor to measure the spatial randomness in point process data and calculated the ANN ratio as the observed mean distance to the nearest neighbor divided by the expected mean distance under the assumption of complete spatial randomness (CSR). In their paper, the

ANN ratio was tested by applying it to a synthetic random distribution and to the distributions of three species of grassland plants. The average nearest neighbor has been used to detect spatial clustering in many applications including analysis of the spatial and spatio-temporal distribution of a vector-borne disease, analysis of public health facilities and the analysis of patterns of spatial distribution of artifacts (Mansour, 2016; Mollalo, Alimohammadi, Shirzadi, & Malek, 2015; Whallon, 1974).

Ripley's K function examines spatial patterns at different scales simultaneously. Given a study area, a researcher can use Ripley's K function to determine if point patterns (e.g. trees) are clustered, dispersed, or randomly distributed throughout the study area. Applications of Ripley's K function can be found in many papers, including the analysis of spatial groupings amongst graves, the analysis of the spatial distribution of diseases and the analysis of spatial distribution patterns of plant communities (Haase, 1995; Mollalo et al., 2015; Sayer & Wienhold, 2013). Unlike ANN and Ripley's K functions, local Moran's I and Getis-Ord Gi* statistics deal with both the locations and the observed values of spatial data. Both statistics are calculated for each feature (point or polygon). Getis-Ord Gi* identifies statistically significant spatial clusters of high values and low values. Local Moran's I identify statistically significant spatial clusters of features with high or low values and spatial outliers. Applications of local Moran's I and Getis-Ord Gi* can be found in many papers including the detection of spatial clusters of upper primary education level in India, identification of pollution hotspots of lead in urban soils of Galway, Ireland, the assessment of the short-term risk of mountain pine beetle and the identification of hot spots on freeways from an incident management database (Bone, Wulder, White, Robertson, &

Nelson, 2013; Jana & Sar, 2016; Songchitruksa & Zeng, 2010; Zhang, Luo, Xu, & Ledwith, 2008).

The ANN ratio and Ripley's K function methods are developed specifically for point process data. However, in practice, these methods are often used on areal data. To our knowledge, the performance of the ANN ratio and Ripley's K function for areal data has never been evaluated. Such an evaluation is desirable, as the null hypothesis of the statistical tests associated with both of these methods is that the locations of the observed data arise from a homogeneous Poisson process, an assumption that is clearly violated for areal data. Local Moran's I and Getis Gi* statistics can be used for both point process and areal data, though slightly different assumptions are needed for each case. In this thesis, we conduct a simulation study to evaluate the performance of each method for areal data under different types of spatial dependence and three different areal structures. Chapter 2 presents the detailed description of four spatial clustering methods. Chapter 3 presents simulation studies. Chapter 4 and 5 are designated for results and discussion.

CHAPTER 2: METHODS

In this section, we review the four spatial clustering detection methods we plan to compare in our simulation study. The first two methods, the average nearest neighbor (ANN) ratio, and Ripley's K function are used to assess spatial patterns in the locations of observed data. The null hypothesis for each of these methods is that the observed locations exhibit complete spatial randomness (CSR), that is, the observation locations arise from a two-dimensional homogeneous Poisson process. Formally, a stochastic process is said to be a homogeneous Poisson process with rate λ if the number of events in any bounded region A, denoted $N(A)$, is Poisson distributed with mean intensity $\lambda|A|$, that is, $\Pr(N(A) = n) = e^{-\lambda|A|}(\lambda|A|)^n/n!$, where $|A|$ denotes the area of A. Given that there are n events in A, those events form an independent random sample from a uniform distribution on A (Cressie, 1994).

The second two methods, local Moran's I statistic and Getis-Ord Gi* statistic are used to assess spatial patterns in the observed data values (rather than patterns in the locations alone). The null hypothesis for these methods is that there is no spatial association in the observed values. An example serves to illuminate the difference between the type of spatial patterns evaluated by the ANN ratio and Ripley's K function versus the type of spatial patterns evaluated by local Moran's I and Getis-Ord Gi* statistics. Suppose we wish to study the spatial patterns of trees in a field. We can assess whether trees are clustered by location using the ANN ratio and/or Ripley's K function. These methods answer the

question ‘do trees tend to be located close to other trees?’. We can assess whether trees are clustered by height using local Moran’s I and/or Getis-Ord Gi* statistics. These methods answer the question ‘do tall (short) trees tend to be located close to other tall (short) trees?’. While ANN and Ripley’s K function are developed specifically for point process data, the development of local Moran’s I and Getis Gi* is more general and accommodates both point process and areal data. When applying local Moran’s I statistic or Getis Gi* statistic, the observation locations are treated as fixed, while the associated values are considered random variables. This framework is different from that used for ANN and Ripley’s K function, in which the observation locations themselves are random.

2.1 AVERAGE NEAREST NEIGHBORS

The average nearest neighbor (ANN) method was developed by Clark and Evans in 1954 (Clark & Evans, 1954) in the context of pattern classification within plant populations. In this context, the observed data consists of locations of observed plants, measured as coordinates in two-dimensional space. This method quantifies the randomness (or lack thereof) among the observed point locations by measuring the distance from each point to its nearest neighbor and using these distances to compute the average nearest neighbor (ANN) ratio given by

$$R = \frac{\bar{r}_O}{\bar{r}_E} \tag{1}$$

where $\bar{r}_O = \frac{\sum_{i=1}^{N} r_i}{N}$, r_i denotes the distance from the i^{th} individual to its nearest neighbor, $\overline{r}_E = \frac{1}{2\sqrt{\rho}}$ is the expected value of $\overline{r}_O$ under CSR for an infinite study area, $\rho = \frac{N}{|A|}$ is the density of the observed distribution, and $|A|$ is the size of the study area. Under CSR, $E(R) = 1$, and for a perfectly clustered distribution (i.e., all points fall at the same

location), $E(R) = 0$. Values of R greater than one indicate that the points are distributed more uniformly than expected under CSR and under the maximum possible spacing (a hexagonal grid) R is 2.1491.

Clark and Evans use the average nearest neighbor framework to develop a hypothesis test to determine if an observed set of point locations exhibits CSR. This test assumes that the distribution of r_O under the null hypothesis of CSR is approximately normal. The z-score for the statistic is calculated by

$$z = \frac{\bar{r}_O - \bar{r}_E}{\sigma_{\bar{r}_E}} \tag{2}$$

and $\sigma_{\bar{r}_E}$ is the standard error of the mean distance to the nearest neighbor under CSR. It can be shown that $\sigma_{\bar{r}_E} = \frac{0.26136}{\sqrt{N\rho}}$. A negative z-score indicates clustering, and a positive score indicates dispersion.

Clark and Evans note some limitations of their procedure. Ideally, the true density of the underlying population should be known, and this might be difficult to measure when N is large. In such situations, the estimated value of mean distance to its nearest neighbor can be calculated from quadrants selected from a random sample (Clark & Evans, 1954). The calculation of $\overline{r}_E$ assumes an infinite study area, which is never the case in practice. The ANN ratio is also sensitive to the chosen study area, as expanding or contracting the study area may alter the distance to the nearest neighbor for some observations. Since the measure is based only on the distance to its nearest neighbor, it fails to distinguish between certain types of spatial dependence (e.g., tightly clustered points in one place vs pairs of points scattered in population). In this situation, Clark and Evans suggest an extension to

this measure by constructing a circle for each observation with an infinite radius, dividing the circle into equal sectors, and measuring the distance from the individual to its nearest neighbor for each of the sectors.

When ANN is applied to large areal units rather than points, only the centroid of each unit is used to calculate the ANN, but it is possible that the individual spatial units are closer to each other and at the same time centroids are distributed uniformly (Clark & Evans, 1954).

2.2 RIPLEY'S K FUNCTION

The ANN ratio discussed in section 1.1 is based on first-order statistics, i.e., the mean of the distances between observation locations. One of the limitations of the method is the inability to test point patterns at different scales simultaneously (Ripley, 1977). For example, it is possible for data to be clustered at a small scale and clustered at a larger scale (i.e., the clusters are clustered) or clustered at a small scale but dispersed at a large scale (i.e., the clusters occur at somewhat regular intervals). Ripley's K function is a second-order spatial analysis tool (i.e., uses variances of the distances between observations) that can address the issue of scale-dependent spatial patterns. Here we only define Ripley's K function for univariate spatial patterns in two dimensions, but it can also be extended for multivariate spatial patterns (ex: comparing spatial patterns of two species). The function K is given by

$$K(t) = \frac{Expected\ number\ of\ additional\ points\ within\ radius\ t\ of\ a\ randomly\ chosen\ point}{\lambda},$$

where λ is the density (number of points per unit area) and can be estimated as $\hat{\lambda} = \frac{N}{|A|}$, where N is the observed number of points and $|A|$is the size of the study area. If points follow a homogenous Poisson process (i.e., exhibit CSR), then $K(t) = \pi t^2$ which is the

area of a circle of radius t. An approximately unbiased estimator for $K(t)$ was proposed by Ripley (Dixon, 2014; Ripley, 1976):

$$\widehat{K}(t) = \hat{\lambda}^{-1} \sum_i \sum_{j \neq i} w_{ij}^{-1} \frac{I(d_{ij} < t)}{N}, \tag{3}$$

where d_{ij} is the distance between the i^{th} and j^{th} points, $I(d_{ij} < t)$ is the indicator function with the value of 1 if $d_{ij} < t$ and 0 otherwise, and w_{ij}^{-1} is a weighting factor associated with locations i and j that corrects for edge effects. Correction for edge effects is required if any distance d_{ij} is greater than the distance between point i and the boundary. Because the points outside the boundary are not included in the calculation of $\widehat{K}(t)$, edge effects can lead to a biased estimator of $K(t)$. Various authors have proposed different edge corrections. One of the most commonly used edge corrections assigns w_{ij} a value of 1 if the circle centered at point i which passes through point j is entirely inside the study area and assigns w_{ij} equal to the proportion of the circumference of the circle that falls in the study area otherwise. Getis and Franklin discussed several edge correction techniques in 1987. For example, if the distance d_{ij} is greater than the distance between point i and the nearest boundary (denoted e, see Figure 2.1), the weighting function is given by (Getis & Franklin, 1987),

$$w_{ij} = 1 - cos^{-1}(e/d_{ij})/\pi \tag{4}$$

To test for CSR, the estimator $\hat{L}(t) = [\widehat{K}(t)/\pi]^{1/2}$ is sometimes used in practice and $E(\hat{L}(t)) = t$ under CSR (Ripley, 1979). If the observed value of K is larger than the expected value of K for a given distance, the distribution is more clustered than CSR at that distance. If the observed value of K is smaller than the expected value of K, the distribution is more dispersed than the random distribution at that distance. See Figure 2.2

for examples of data exhibiting clustering, dispersion, and CSR. A plot of $\widehat{K}(t)$ versus πt^2 may reveal the deviations from the expected value under CSR. Statistical significance of the deviation can be tested as $\widehat{K}(t) - \pi t^2 = 0$ at each distance t. Usually, the distribution of $\widehat{K}(t)$ is simulated under the null hypothesis of CSR, and critical values from the simulated distribution are used to define the rejection region.

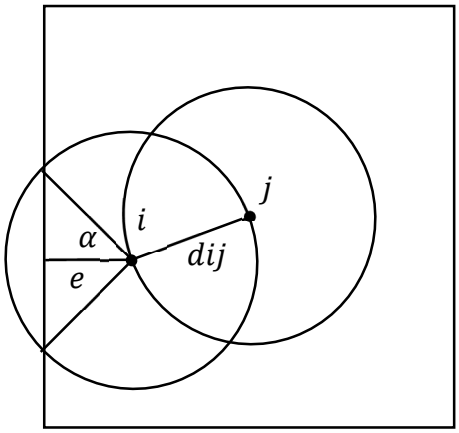

Figure 2.1 Example of weighted edge correction, d_{ij} is the distance between the i^{th} and j^{th} points, e is the distance between point i and the nearest boundary, the value of α gives the inverse cosine function in Eq. 4 (Haase, 1995).

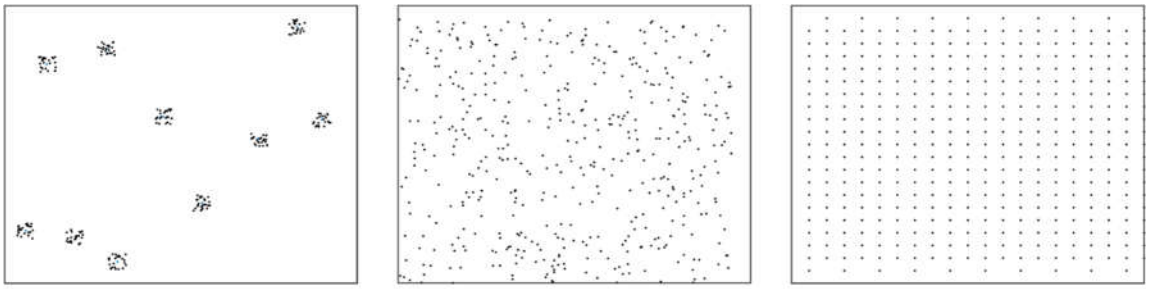

Figure 2.2 Examples of spatial patterns of clustering (left), CSR (middle), and dispersion (right).

2.3 LOCAL MORAN'S I

In this section and the next, we consider two local spatial statistics: Local Moran's I and Getis-Ord Gi* which assess spatial dependence in the values of the observed data, rather than assessing patterns in the observation's locations. The null hypothesis for these methods is that there is no spatial association in the observed values. Local Moran's I statistic is a decomposition of global Moran's I statistic which measures local spatial autocorrelation. Global statistics might not sufficiently capture the spatial patterns in large data sets if the spatial association is not constant over the entire study area. As discussed by Anselin (Anselin, 1995), local indicators of spatial associations (LISA) quantify two distinct forms of spatial dependence by identifying local spatial clusters (hot spots) and outliers. The local Moran's I statistic associated with the ith observed data point is given by

$$I_i = \frac{z_i - \bar{z}}{\sigma^2} \sum_{j=1, j \neq i}^{N} [w_{ij}(z_j - \bar{z})], \tag{5}$$

where z_i is the observed value at location i, $\bar{z}$ is the corresponding sample mean, N, is the number of observations, σ^2 is the variance of the variable z (in practice, σ^2 is usually replaced by the sample variance) and w_{ij} is a spatial weight associated with observations i and j. The spatial weights may be chosen in several ways. For example, we may take $w_{ij} = 1$ if observations i and j are within a pre-specified distance of each other and $w_{ij} = 0$ otherwise. Alternately, the w_{ij} can also be defined using the adjacency of units or as the inverse of the distance between observations I and j or the inverse of the distance squared.

Under the null hypothesis of no spatial association, the expected value of I is given by

$$E[I_i] = -\frac{\sum_{j=1, j\neq i}^{N} w_{ij}}{N-1}.$$

A high positive value of local Moran's I indicates that the location under study has neighbors with similarly high or low values and the location is a part of a cluster. A high negative value of local Moran's I indicates that the location has neighbors with dissimilar values and the location is an outlier. Figure 2.3 illustrates the type of spatial clusters and outliers; high-high (high values surrounded by high values), low-low (low values surrounded by low values), high-low (high values surrounded by low values), and low-high (low values surrounded by high values) (Zhang et al., 2008).

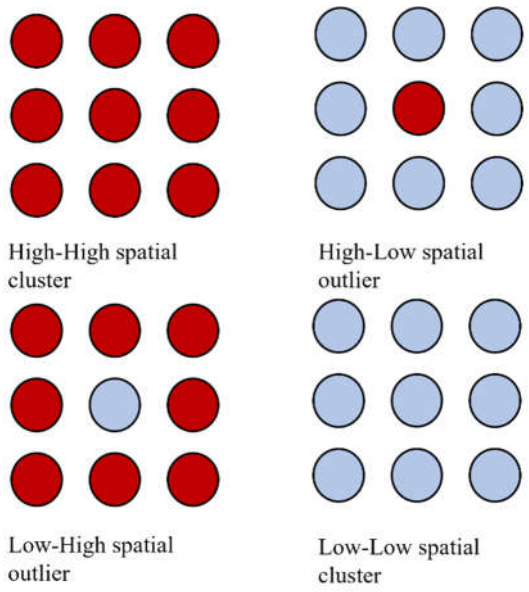

Figure 2.3 Cluster/outlier types. Red indicates high values and blue indicates low values.

Since the exact probability distribution of local Moran's I is hard to obtain, an alternative method called "conditional permutation" is often used to determine how likely the observed spatial distribution of the values is under the null hypothesis of no spatial association (Anselin, 1995). The conditional permutation method proceeds by fixing a

location and randomly permuting the remaining values on all other locations. This process is repeated many times, and local Moran's I value is calculated each time. The significance level is calculated by comparing the actual values with the simulated. A pseudo p-value is calculated by determining the proportion of local Moran's I values obtained from permutations that have higher (for positive local Moran's I) or lower values (for negative local Moran's I) than the observed value. A pseudo value less than 0.05 shows a significant spatial pattern in the data.

2.4 GETIS-ORD GI*

Local Moran's I statistic described cannot discriminate between hot spots (i.e., high-high clustering) and cold spots (low-low clustering). The Getis-Ord Gi* statistic, developed by Getis and Ord (Getis & Ord, 2010), is a local statistic that can distinguish between hot and cold spots. However, the Getis-Ord Gi* statistic cannot identify outliers, so Getis and Ord suggest that their statistic should be used in conjunction with the local Moran's I statistic to understand spatial patterns more fully.

Consider an area subdivided into N regions, $N = 1,2, \dots N$ where each region has a centroid with x and y coordinates in a Cartesian plane. We observe data $z_1, z_2, \dots, z_N$, where z_i is associated with the i^{th} region. The Getis-Ord local statistic is given by (Getis & Ord, 2010)

$$G_i^*(d) = \frac{\sum_{j=1}^{N} w_{ij} z_j}{\sum_{j=1}^{N} z_j}, \tag{6}$$

where w_{ij} is the spatial weight between observations i and j. Usually, w_{ij} is a binary with $w_{ij} = 1$ if observations j is within a threshold distance d of observation i and 0 otherwise. Since the denominator is constant in Eq. 6, a higher value of the set of observations z_j for

which $w_{ij} > 0$ correspond to higher values of G_i^*. The expectation of the G_i^* is given by (Songchitruksa & Zeng, 2010)

$$E(G_i^*) = \frac{W_i}{N},$$

and the variance is given by

$$Var(G_i^*) = \frac{s^2}{N}\frac{W_i(N-W_i)}{(N-1)},$$

where $W_i = \frac{\sum_{j=1}^{N} w_{ij}}{N}$ is the sum of the weights, $\bar{z} = \frac{\sum_{j=1}^{N} z_j}{N}$ and $s^2 = \frac{\sum_{j=1}^{N} z_j^2}{N} - \bar{z}^2$ are the sample mean and sample variance of the random variable Z. Under exact or asymptotical normal conditions, a z-score can be computed via

$$Z(G_i^*) = \frac{\sum_{j=1}^{N} w_{ij} z_j - \bar{z}\sum_{j=1}^{N} w_{ij}^2}{s\sqrt{\frac{N\sum_{j=1}^{N} w_{ij}^2 - (\sum_{j=1}^{N} w_{ij})^2}{N-1}}}.$$

Positive values with higher z-score values indicate clustering of high values (hot spots) and negative values with lower z-score values indicate clustering of low values (cold spots). Getis and Ord explore a Bonferroni correction to correct for multiple testing when computing the Gi* statistic for all N observations. However, this can be unduly conservative when N is large (Getis & Ord, 2010). While the Getis-Ord Gi* statistic was originally developed for areal data, it can also be applied to point process data (Baruch-Mordo, Breck, Wilson, & Theobald, 2008; Siebeneck, Medina, Yamada, & Hepner, 2009; Sokal & Thomson, 2006).

CHAPTER 3: SIMULATIONS

A simulation study is carried out to assess the performance of the four spatial pattern detection methods described in the previous section under different scenarios. In this study, we consider three areal unit structures: structure A_1 is a 20×20 regular grid where the units are of the same size, structure A_2 is the United States (US) counties in the states of North Carolina, Tennessee, South Carolina, Georgia, Alabama, and Mississippi where the units (counties) are roughly the same size but irregularly shaped, and structure A_3 is the Canadian forward sortation areas (FSAs) in the provinces of Alberta, Saskatchewan, and Manitoba where the units have vastly different sizes. Areal structures A_1, A_2, and A_3 contain $n_1 = 400$, $n_2 = 549$, and $n_3 = 267$ areal units, respectively. We generate data under null and alternative hypotheses and the specific data generation mechanisms (DGMs) depend on the method under consideration. For each areal structure and each DGM, 500 datasets are simulated.

Recall that the ANN and Ripley's K methods test for spatial patterns based on the locations of the observations only. These methods are often applied to areal data to determine if areal units having some characteristic of interest exhibit a spatial pattern. For example, researchers might be interested in which counties have experienced a case of a rare disease. We refer to the units of interest as the 'observed' units. Generating observed units on areal structure A_a consists of selecting the observed units from among the n_a total units present in A_a. We consider two sample sizes for each areal structure: $N = \lfloor \frac{n_a}{10} \rfloor$ and

$N = \lfloor \frac{n_a}{4} \rfloor$, and generate data under the null hypothesis of no spatial pattern and under two alternative hypotheses: a single large cluster and multiple smaller clusters. We now describe the data generation mechanism for each of these hypotheses in detail. Under DGM D_1, which corresponds to the null hypothesis, the observed units are selected via uniform sampling without replacement from among the $n_a$total units. Under DGM D_2, which corresponds to the single cluster, units are sampled without replacement and the units in a pre-selected region of the study area are sampled with 10 times higher probability than the units in the rest of the study area. DGM D_3, which corresponds to multiple clusters, is an iterative process. First, an areal unit is selected via uniform random sampling, and this unit is observed, along with units which are directly adjacent to it. Another areal unit is then be sampled without replacement from the remaining unobserved units and is observed along with all adjacent units. This process is continued until N units have been observed. Examples of data generated on each areal structure under DGMs D_1, D_2 and D_3 are shown in Figure 3.1, Figure 3.2, and Figure 3.3.

Recall that local Moran's I and Getis-Ord Gi* test for spatial patterns based on observed data values (rather than based on observed data locations). Typically, these methods are applied to areal data only when each areal unit is associated with some numeric value and researchers wish to determine if these observed values exhibit spatial patterns (such as large values clustering near other values). For each areal structure A_a, we generate N observed values, one value for each unit, under the null hypothesis of no spatial pattern and under three alternative hypotheses: high-high clustering, a mixture of high-high and low-low clustering, and a mixture of high-high clustering and high-low outliers. To generate data under the null hypothesis, (DGM D_4), observed values are independently

generated from a $N(0,1)$ distribution. To generate data exhibiting high-high clustering (DGM D_5), 4 units are selected via uniform sampling without replacement. Observed values for these units and all units adjacent to them are independently generated from a $N\left(5,\frac{1}{2}\right)$ distribution. Observed values for the remaining units are independently generated

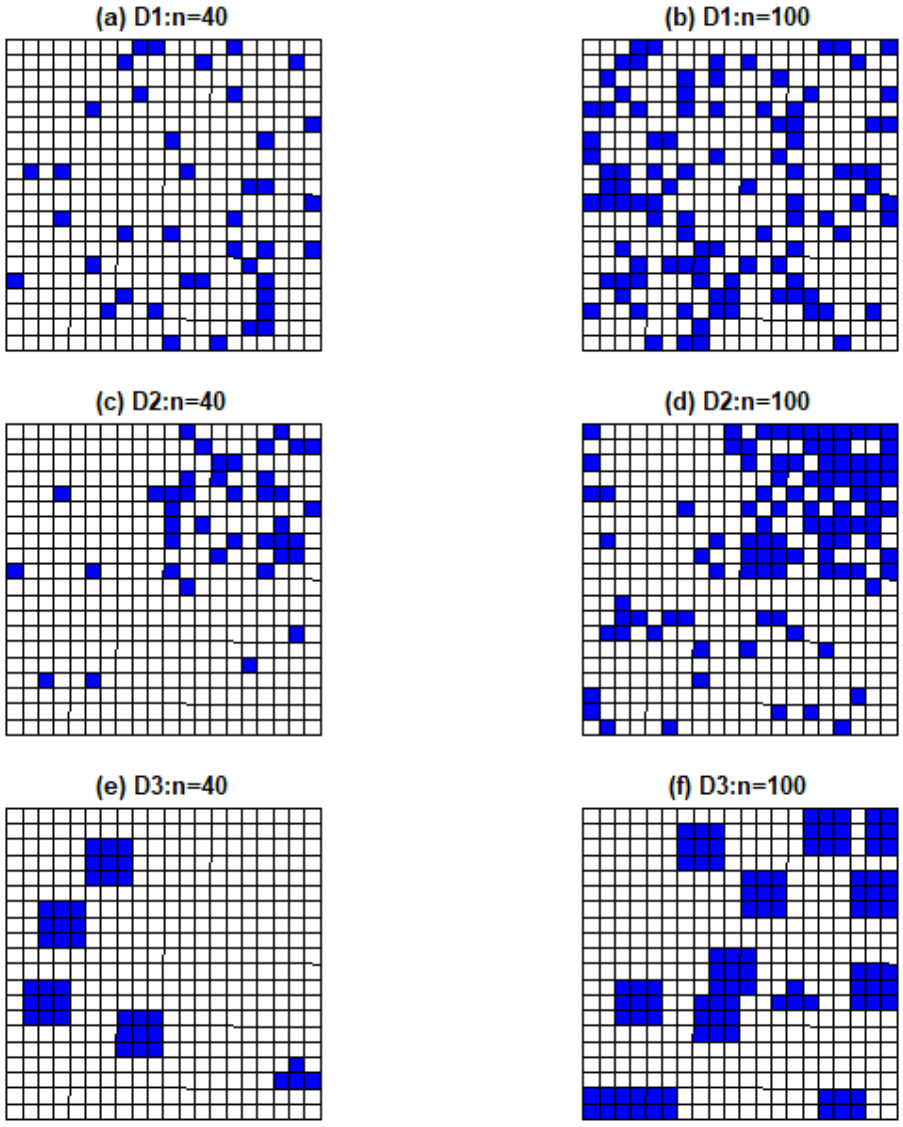

Figure 3.1 Observed units for 20x20 regular grid (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes.

from a $N(0,1)$ distribution. To generate data exhibiting a mixture of high-high and low-low clustering (DGM D_6), 2 units are selected via uniform sampling without replacement.

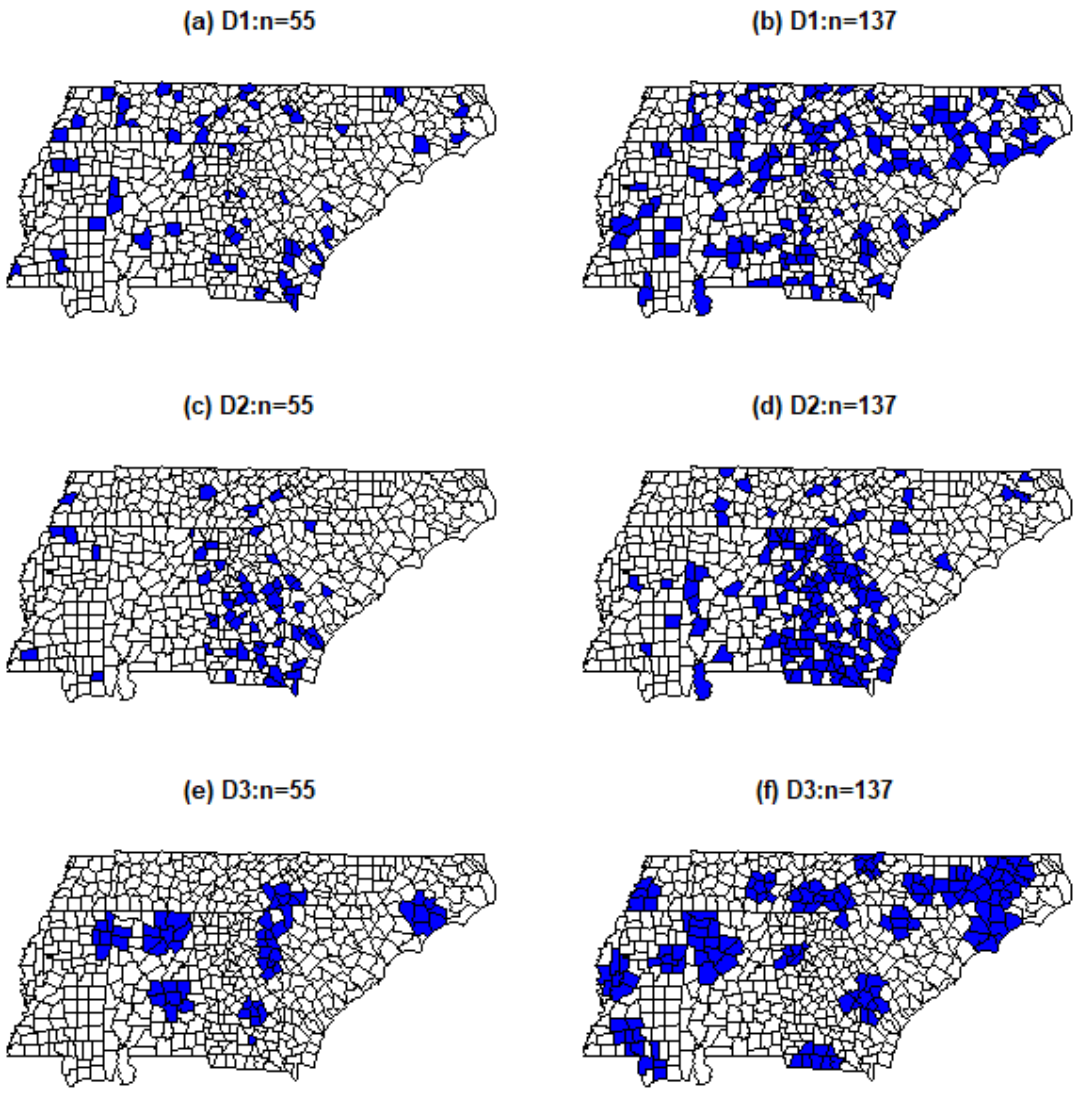

Figure 3.2 Observed units for USA counties (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes.

Observed values for these units and all units adjacent to them are independently generated from a $N\left(5,\frac{1}{2}\right)$ distribution. An additional 2 units are selected via uniform sampling from the units which have not yet been assigned values and are not adjacent to any unit which has already been assigned a value. Observed values for these units and all units adjacent to

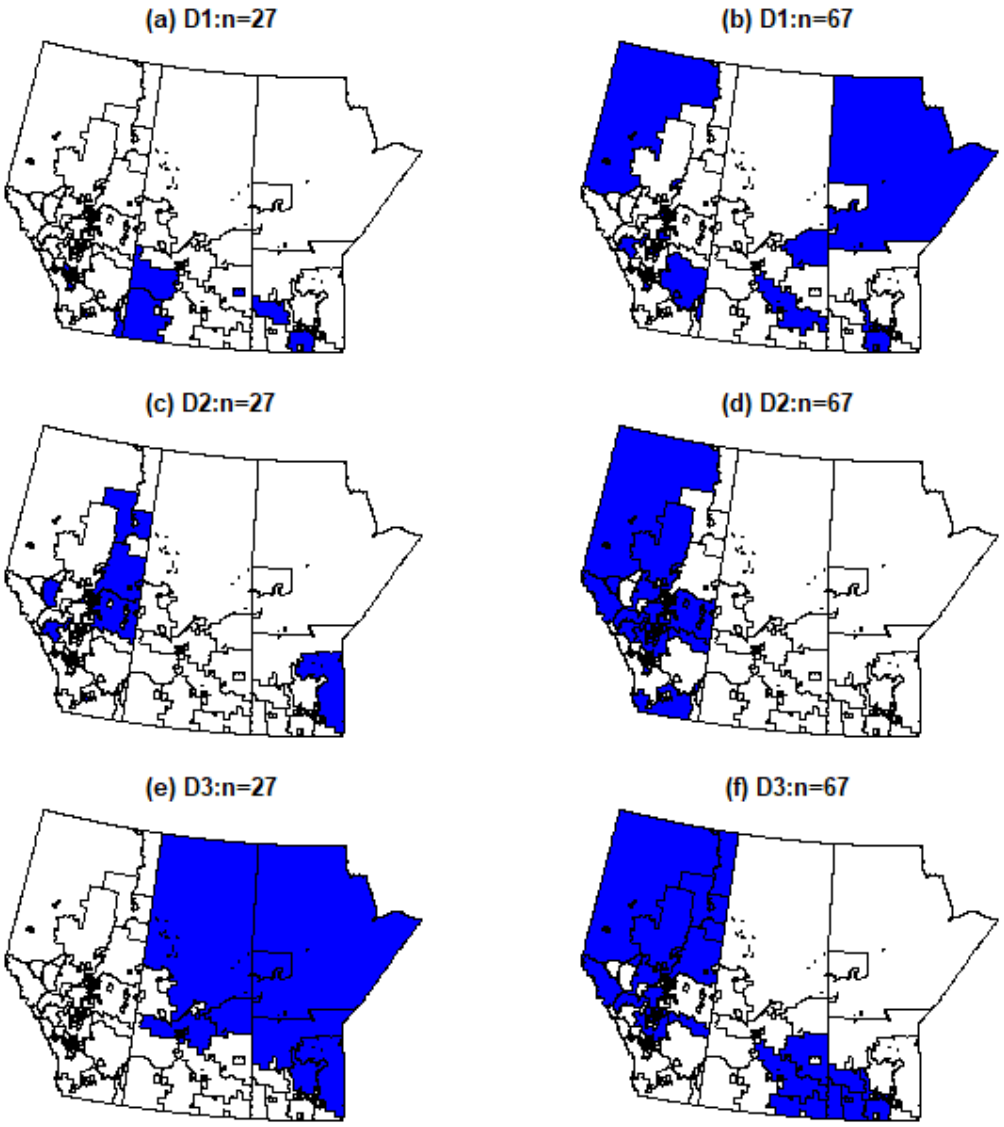

Figure 3.3 Observed units for CA FSAs (blue) for a one simulation under three DGMs (D1: CSR, D2: Single cluster, D3: Multiple cluster) for two sample sizes.

them are independently generated from a $N\left(-5, \frac{1}{2}\right)$ distribution. Observed values for the remaining units are independently generated from a $N(0,1)$ distribution. To generate data exhibiting a mixture of high-high clustering and high-low outliers (DGM D_7), 2 units are selected via uniform sampling without replacement. Observed values for these units and

all units adjacent to them are independently generated from a $N\left(5,\frac{1}{2}\right)$ distribution. An additional 2 units are selected via uniform sampling without replacement from among the

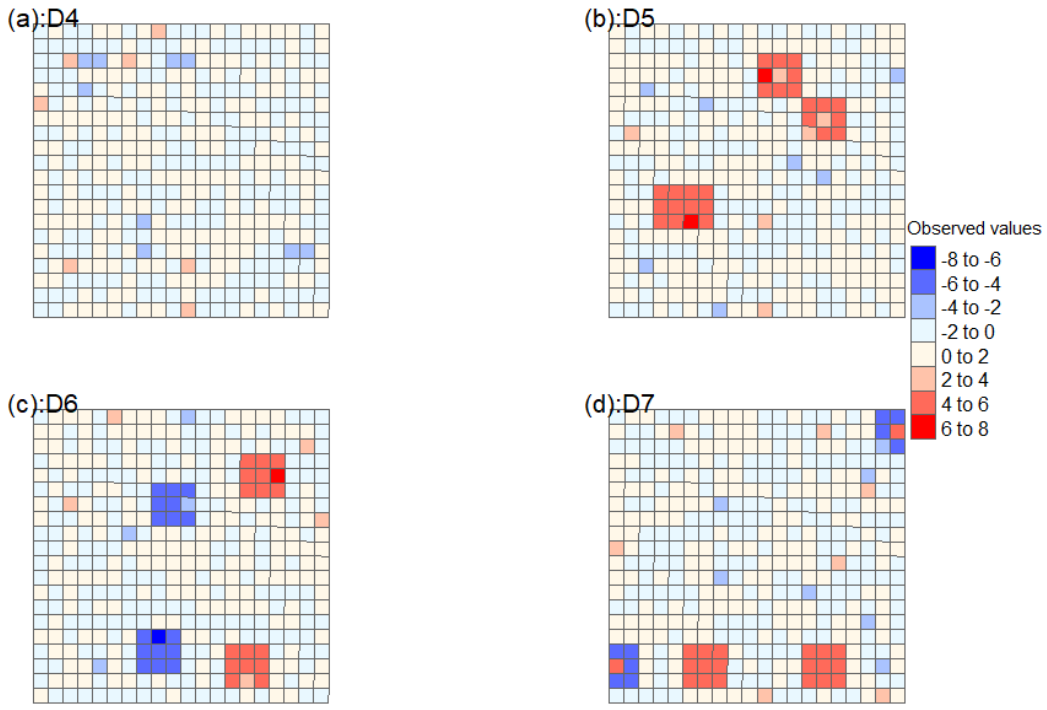

Figure 3.4 Observed values generated for 20x20 regular grid under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers.

units which have not already been assigned a value and are not adjacent to units which have already been assigned a value. Observed values for these units are independently generated from a $N\left(5,\frac{1}{2}\right)$ distribution, and values for all units adjacent to them are independently generated from a $N\left(-5,\frac{1}{2}\right)$ distribution. Observed values for the remaining units are independently generated from a $N(0,1)$ distribution. Figure 3.4, Figure 3.5, and Figure 3.6 show examples of data generated on each areal structure under each mechanism. For each data set generated under DGM $D_1 - D_3$ ANN, and Ripley's K function are calculated by using the centroids of the observed units as the observation locations. ANN

and z-score is calculated manually using Eq.1 and Eq.2 and for this we use one observation window which is the whole study area for all the simulations. ANN is also calculated for each simulated dataset using the R function *nni* which also computes a z-score and for this we use a different observation window, a rectangle that encloses only the centroids selected for each simulation. For DGM D_1 we perform an $\alpha = 0.05$ level two-tailed test by rejecting the null hypothesis if the absolute value of the z-score exceeds the 0.975 quantile of a standard normal distribution. For DGMs D_2 and D_3, we perform a left-tailed test (indicative of clustering) by rejecting the null hypothesis if the z-score falls below the 0.05 quantile of a standard normal distribution.

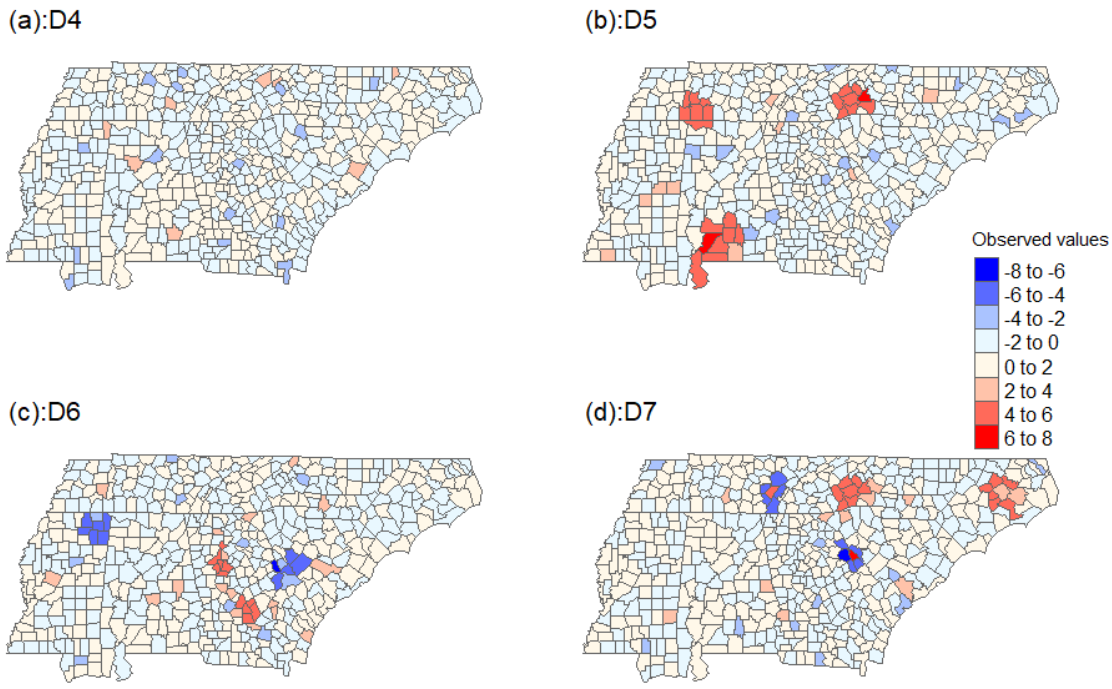

Figure 3.5 Observed values generated for USA counties under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers.

Recall that computing Ripley's K function (Eq. 3) requires specifying a radius at which clustering is evaluated. For each areal structure, a sequence of 5 radii are evaluated,

with the smallest radius being equal to the twice the smallest distance between any two-unit centroids and the largest radius being equal to one-quarter of the width of the study area. For each dataset, Ripley's K function is evaluated at each radius using the Kest function in R with the correction input as "Ripley", which implements the edge correction for rectangular and polygonal windows. To reduce the computational time, we approximated the distribution of Ripley's K function at each radius is approximated via the Monte Carlo method with 1000 replications, and critical values are approximated with quantiles of the Monte Carlo samples at each radius once for all the simulations.

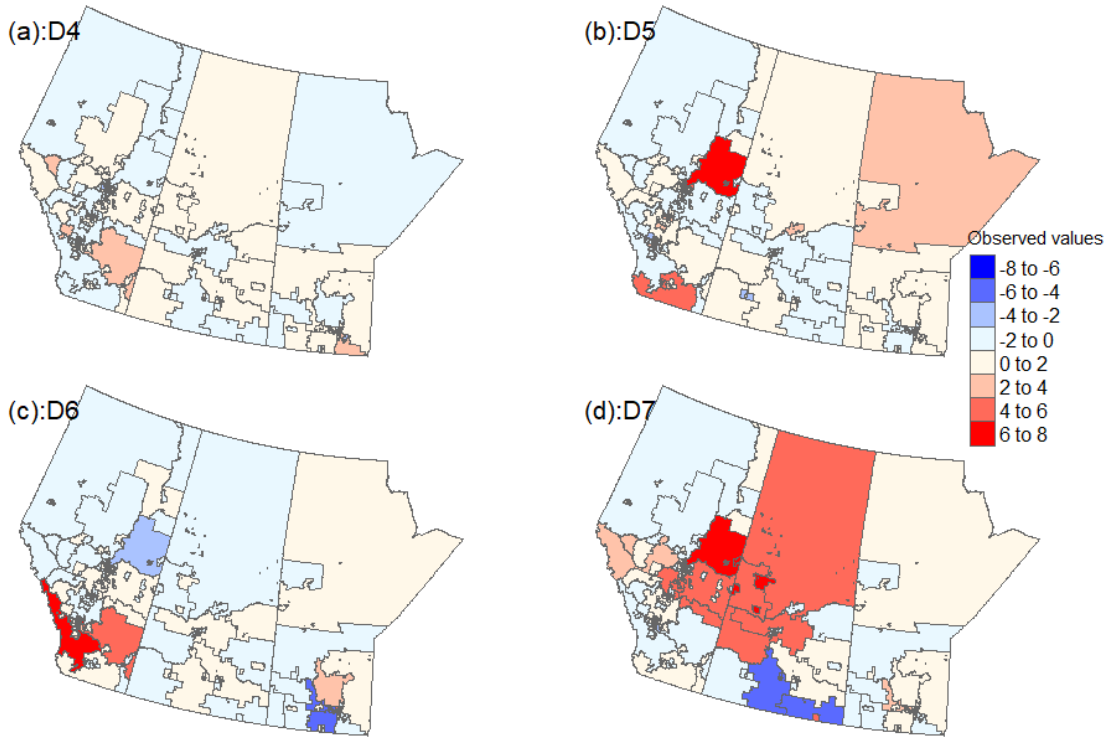

Figure 3.6 Observed values generated for CA FSAs under 4 DGMs for a one simulation; D4: No spatial pattern, D5: high-high clustering, D6: high-high and low-low clusters, D7: high-high clusters and high-low outliers.

For DGM D_1 we perform an $\alpha = 0.05$ level two-tailed test by rejecting the null hypothesis

if the absolute value of the K estimate exceeds the 0.975 quantile of the Monte Carlo samples. For DGMs D_2 and D_3, we perform a right-tailed test (indicative of clustering) by rejecting the null hypothesis if the K estimate exceeds the 0.95 quantile of the Monte Carlo samples.

We compute local Moran's I and Getis-Ord Gi* statistics for each unit in each data set simulated under DGM $D_4 - D_7$. For both methods, weights are assigned based on the adjacency of the units; neighboring units are given weight 1 while others are given weight 0. The distribution of local Moran's I at each unit is approximated using the conditional permutation approach explained in Chapter 2. Local Moran's I is calculated using *localmoran_perm* function in R with input p.adjust.method equals to "holm". This approach applies the Holm-Bonferroni correction for multiple comparisons, which is uniformly more powerful than the Bonferonni correction and remains valid even in the presence of dependence among the test statistics (Holm, 1979). For each DGM, we apply a two-tailed test and any unit with a significant p-value (<0.05) is classified as a part of a cluster (positive local Moran's I value) or an outlier (negative local Moran's I value) based on the sign of the associated test statistic.

Getis-Ord Gi* is calculated as a z-score using the localG function in R. The Holm-Bonferroni correction is used to correct for the multiple testing. For each DGM $D_4 - D_7$, we apply a two-tailed test and any unit with a significant p-value (<0.05) is classified as a part of a high-high cluster or low-low cluster based on the sign of the associated test statistic.

We assess the performance of ANN and Ripley's K via empirical type I error rate and empirical power. For simulations under the null hypothesis (D_1), we report the global

type I error rate for ANN and the type I error rate at each radius for Ripley's K function. For simulations under alternative hypotheses (D_2 and D_3), we report the global power for ANN and the power at each radius for Ripley's K function.

We evaluate the performance of local Moran's I and Getis-Ord Gi* by calculating the classification accuracy under each DGM and summarizing the results using maps for each of these methods. Local Moran's I and Getis-Ord Gi* classify units into 3 categories (local Moran's I classifies units as part of a cluster, an outlier, or neither, Getis- Ord Gi* classifies units as part of a high-high cluster, part of a low-low cluster, or neither). Finally, we report average classification accuracies over all units.

CHAPTER 4: RESULTS

4.1 ANN AND RIPLEY'S K

Table 4.1 summarizes the empirical type I error rate (i.e., empirical probability of rejecting the null hypothesis of CSR when units are generated under CSR) and empirical power (i.e., the empirical probability of rejecting the null hypothesis of CSR when units are clustered) of ANN for the three areal structures and two sample sizes. According to Table 4.1, the type I error of all three areal structures is greater than 0.05 and its highest (0.99) for Canadian (CA) FSAs where the units are vastly different sizes. When the sample size increases, the type I error of each areal structure also increases. Under the alternative hypothesis, the power of detecting a single cluster is 1.00 for CA FSAs which is greater than both the regular grid and USA counties. When the sample size increases, the power of detecting a single cluster decreases to 0 for regular grid and USA counties.

Table 4.1 Type I error and power of ANN. $n_a = 267$ for Canadian FSAs, $n_a = 400$ for regular grid, $n_a = 549$ for USA counties.

DGM (Data generation mechanism)	Quantity	N	Canadian FSAs	Regular 20 x 20 grid	USA counties
H_0: CSR: D1	Type I error	$\lfloor n_a/4 \rfloor$	0.99	0.39	0.35
		$\lfloor n_a/10 \rfloor$	1.00	1.00	0.99
H_a: Single Cluster: D2	Power	$\lfloor n_a/4 \rfloor$	1.00	0.04	0.27
		$\lfloor n_a/10 \rfloor$	1.00	0.00	0.00
H_a: Multiple Clusters: D2	Power	$\lfloor n_a/4 \rfloor$	1.00	1.00	1.00
		$\lfloor n_a/10 \rfloor$	1.00	0.00	0.95

The results in Table 4.1 are calculated using the centroids of areal units and the observation window is specified using the whole study area. When we calculate the results

in Table 4.2, we use a different observation window, a rectangle that encloses only the centroids selected for each simulation. If we compare the two tables, we see that in Table 4.1, the type I error is noticeably different, but that there is no clear pattern of increase or decrease. The power in table 4.2 is lower for all scenarios except the USA counties under the larger sample size.

Table 4.2 Type I error and power of ANN with different windows for each simulation of data. $n_a = 267$ for Canadian FSAs, $n_a = 400$ for regular grid, $n_a = 549$ for USA counties.

DGM (Data generation mechanism)	Quantity	N	Canadian FSAs	Regular 20 x 20 grid	USA counties
H_0: CSR: D1	Type I error	$\lfloor n_a/4 \rfloor$	0.83	0.75	0.06
		$\lfloor n_a/10 \rfloor$	1.00	1.00	0.01
H_a: Single Cluster: D2	Power	$\lfloor n_a/4 \rfloor$	0.76	0.00	0.24
		$\lfloor n_a/10 \rfloor$	1.00	0.00	0.36
H_a: Multiple Clusters: D2	Power	$\lfloor n_a/4 \rfloor$	0.93	0.53	0.99
		$\lfloor n_a/10 \rfloor$	1.00	0.00	0.96

Table 4.3 Type I error and power of Ripley's K for 5 different radii (20x20 regular grid).

DGM (Data generation mechanism)	Type I error/ Power									
	R1=2040		R2=2809		R3=3578		R4=4346		R5=5115	
	N=40	N=100	N=40	N=100	N=40	N=100	N=40	N=100	N=40	N=100
H_o: CSR: D1	0.98	1.00	0.92	1.00	0.84	0.78	0.75	0.49	0.90	1.00
H_a: Single cluster: D2	0.88	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H_a: Multiple clusters: D3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.79

Table 4.3, Table 4.4, and Table 4.5 summarize the type I error and the power of Ripley's K function under the same three data generation mechanisms for the regular grid, USA counties, and CA FSAs, respectively, for different sample sizes and 5 different radii. The type I error rate is far above its nominal level in almost all cases. For USA counties

(Table 4.4), the power is high (closer to 1) for all radii except the smallest radius R1. For CA FSAs (Table 4.5), the power is 1.0 for all radii except for R1. In general, the power for detecting clusters is high in all three areal structures under all radii except for the minimum radius R1 under USA counties and CA FSAs.

Table 4.4 Type I error and power of Ripley's K for 5 different radii (USA counties).

DGM (Data generation mechanism)	Type I error/ Power									
	R1=22935		R2=72327		R3=121720		R4=171112		R5=220505	
	N=55	N=137	N= 55	N=137	N= 55	N=137	N= 55	N=137	N= 55	N=137
H_o: CSR: D1	0.98	1.00	0.85	0.62	0.85	0.64	0.77	0.73	0.79	0.71
H_a: Single cluster: D2	0.19	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H_a: Multiple clusters: D3	0.35	0.00	1.00	1.00	1.00	1.00	1.00	0.98	0.94	0.91

Table 4.5 Type I error and power of Ripley's K for 5 different radii (CA FSAs).

DGM (Data generation mechanism)	Type I error/ Power									
	R1=1344		R2=90671		R3=179999		R4=269326		R5=358653	
	N=27	N=67	N=27	N=67	N=27	N=67	N=27	N=67	N=27	N=67
H_o: CSR: D1	0.06	0.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H_a: Single cluster: D2	0.08	0.38	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H_a: Multiple clusters: D3	0.25	0.49	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

4.2 LOCAL MORAN'S I AND GETIS-ORD GI*

We calculate the type I error (number of times units are incorrectly classified as a part of cluster or outlier when they are not) for both local Moran's I (Figure 4.1, Figure 4.3, Figure 4.5) and Getis-Ord Gi* (Figure 4.2, Figure 4.4, Figure 4.6) under the four DGMs for all three areal structures. For local Moran's I, the type I error under the null hypothesis (Figure

4.1a, Figure 4.3a, Figure 4.5a) is below 0.05 for all the areal structures. and the type I error is below 0.05 for all the DGMs for USA counties. There are few units where the type I errors are between 0.05 and 0.2 under the two scenarios of the mixture of high-high and low-low clusters (Figure 4.1c, Figure 4.5c) and high-high clusters and high-low outliers (Figure 4.1d, Figure 4.5d). For Getis-Ord Gi*, the type I error rate under DGM 1 is less than 0.05 for each areal structure (Figure 4.2a, Figure 4.4a, Figure 4.6a). The type I error rate for Getis-Ord Gi* under the scenario of high-high clustering is below 0.05 for most units, but between 0.05 and 0.2 for a small number of units, most noticeably in the FSAs (Figure 4.6b). Under the two scenarios of the mixture of high-high and low-low clusters (Figure 4.2c, Figure 4.4c) and high-high clusters and high-low outliers (Figure 4.2d, Figure 4.4d), the type I error rate of Getis-Ord Gi* is slightly inflated (between 0.05 and 0.2) for the regular grid and USA counties and the numbers are more inflated (between 0.4 and 0.6) for the CA FSAs (Figure 4.6c and Figure 4.6d).

Figure 4.7, Figure 4.8, and Figure 4.9 present the classification accuracy (number of times units are correctly classified as a part of a cluster or outlier) of local Moran's I under the DGM 5, 6, and 7 for each areal structure. Accuracy of detecting clusters under the scenarios of high-high clusters (Figure 4.7a, Figure 4.8a), the mixture of high-high and low-low clusters (Figure 4.7b, Figure 4.8b), and the mixture of high-high clusters and high-low outliers (Figure 4.7c, Figure 4.8c) is above 0.6 for most of the units in regular grid and USA counties. For the CA FSAs, the accuracy of cluster detection ranges from 0.05 to 1 (Figure 4.9a, Figure 4.9b, Figure 4.9c) and most of the units which are smaller compared to other units have accuracy above 0.8 while few units have an accuracy between 0.05 and

0.2. The power of detecting outliers is above 0.8 for all three areal structures (Figure 4.7d, Figure 4.8d, Figure 4.9d).

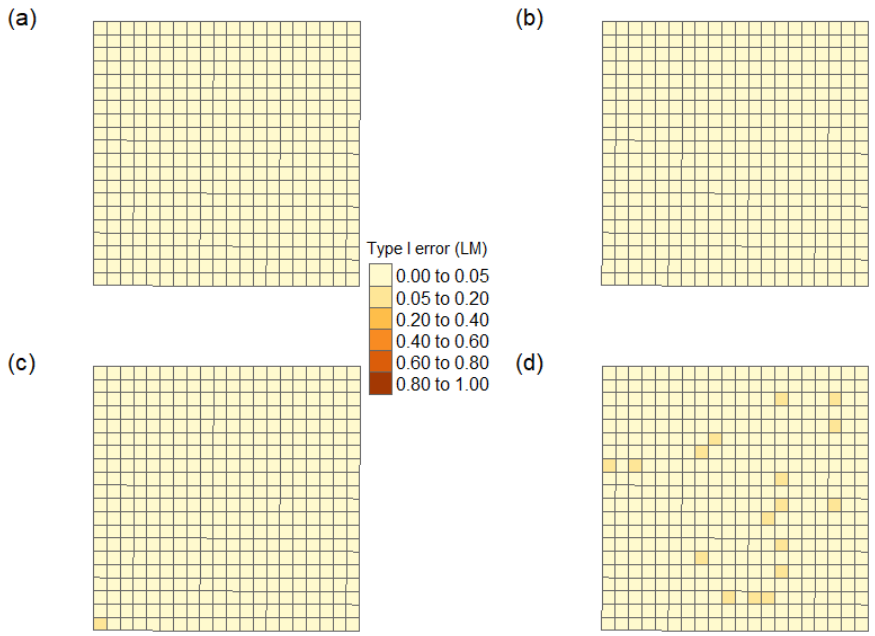

Figure 4.1 Type I error of Local Moran's I (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

Figure 4.10, Figure 4.11, and Figure 4.12 present the classification accuracy (number of times units are correctly classified as a part of a high-high cluster or low-low cluster) of Getis-Ord Gi* under DGMs 5, 6 and 7 for each areal structure. All the units in the regular grid and USA counties have a classification accuracy greater than 0.8. The majority of the units in CA FSAs which are smaller compared to other units have accuracy above 0.8 and the rest of the units have an accuracy between 0.05 and 0.08.

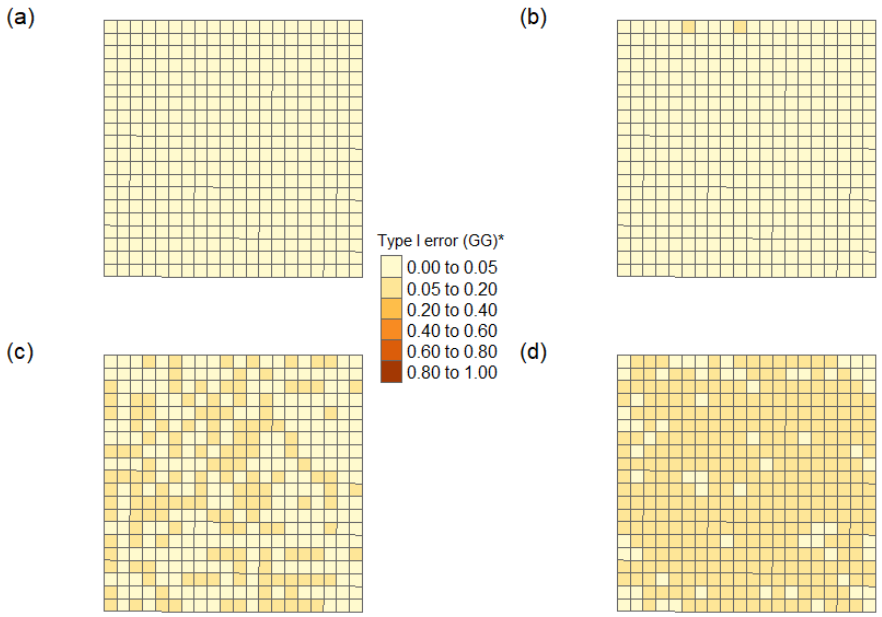

Figure 4.2 Type I error of Getis-Ord Gi* (20x20 regular grid) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

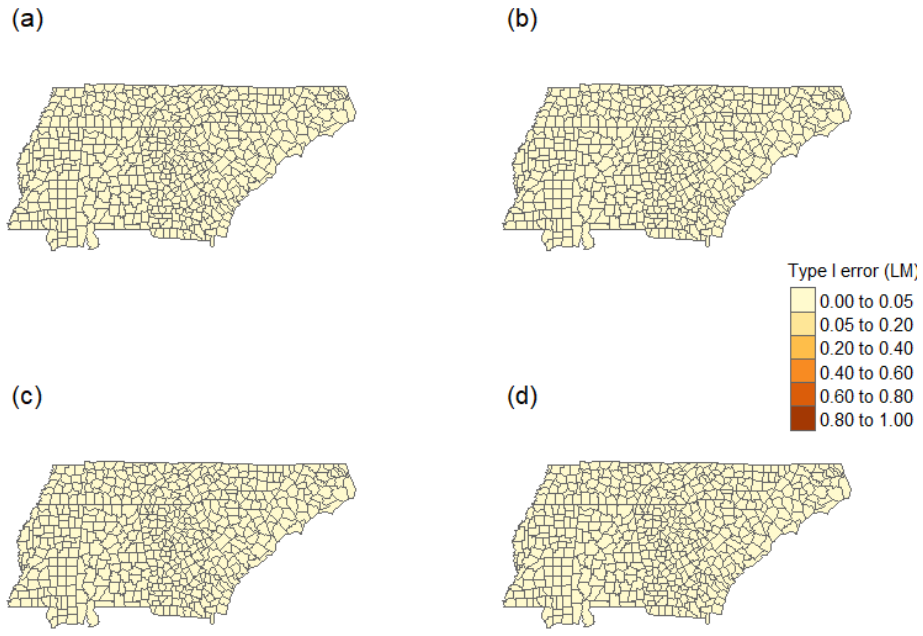

Figure 4.3 Type I error of Local Moran's I (USA counties) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high

and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

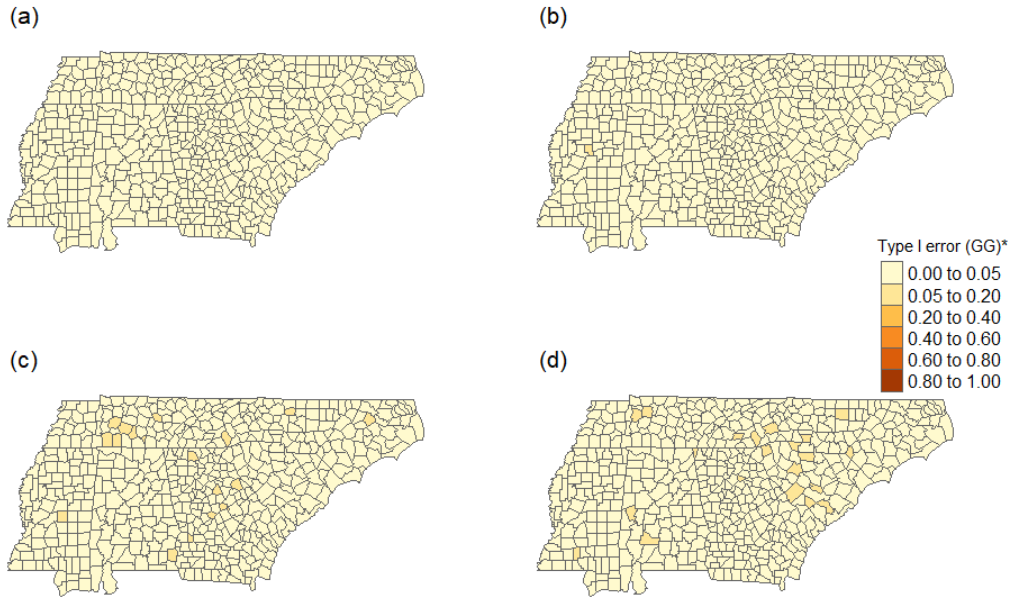

Figure 4.4 Type I error of Getis-Ord Gi* (USA counites) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

Figure 4.5 Type I error of Local Moran's I (CA FSAs) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

Figure 4.6 Type I error of Getis-Ord Gi* (CA FSAs) (a) Under the null hypothesis of no spatial pattern; (b) Under the scenario of high-high clusters; (c) Under the scenario of the mixture of high-high and low-low clusters; (d) Under the scenario of the mixture of high-high clusters and high-low outliers.

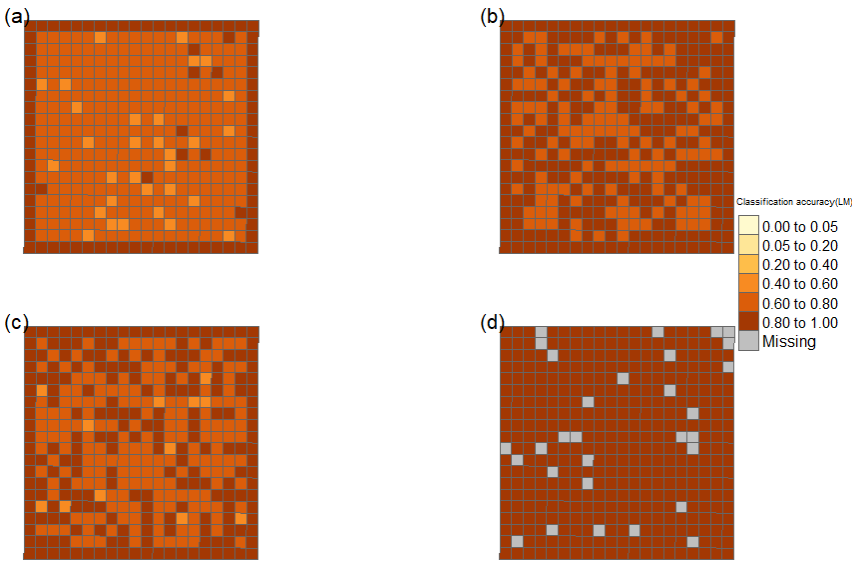

Figure 4.7 Classification accuracy of local Moran's I (20x20 regular grid) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers.

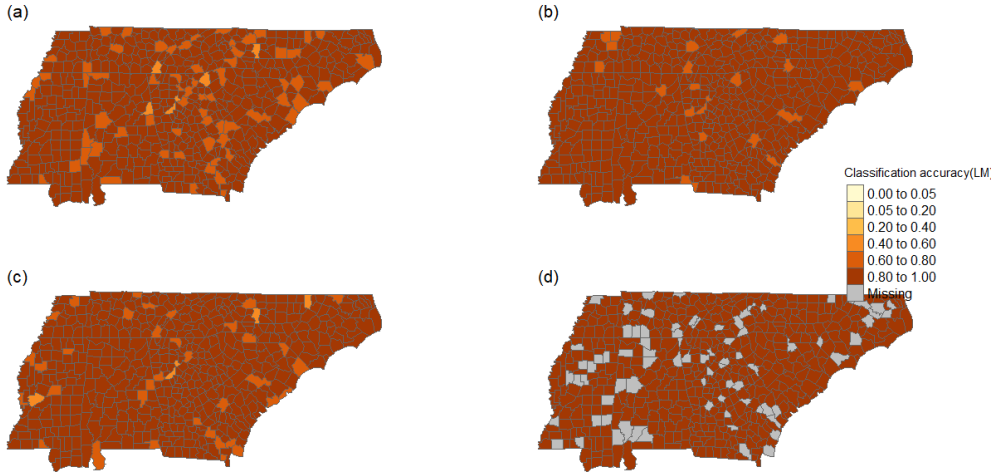

Figure 4.8 Classification accuracy of local Moran's I (USA counties) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers.

Figure 4.9 Classification accuracy of local Moran's I (CA FSAs) (a) Detecting clusters under the scenario of high-high clusters; (b) Detecting clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting clusters under the scenario of the mixture of high-high clusters and high-low outliers; (d) Detecting outliers under the scenario of the mixture of high-high clusters and high-low outliers.

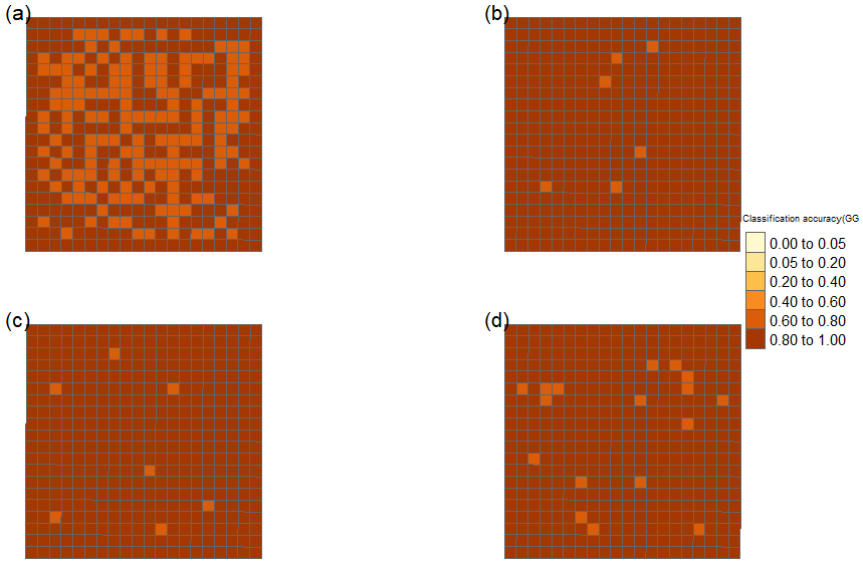

Figure 4.10 Classification accuracy of Getis-Ord Gi* (20x20 regular grid) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers.

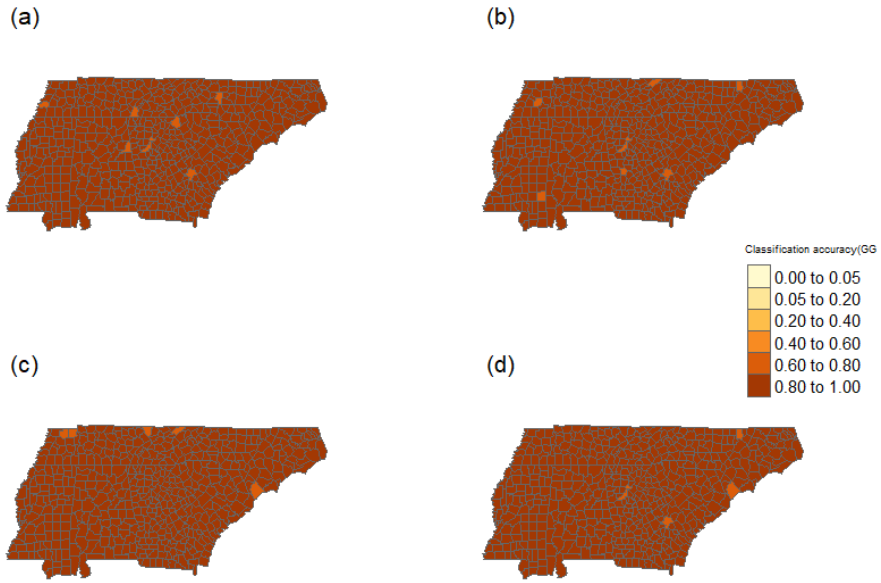

Figure 4.11 Classification accuracy of Getis-Ord Gi* (USA counites) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers.

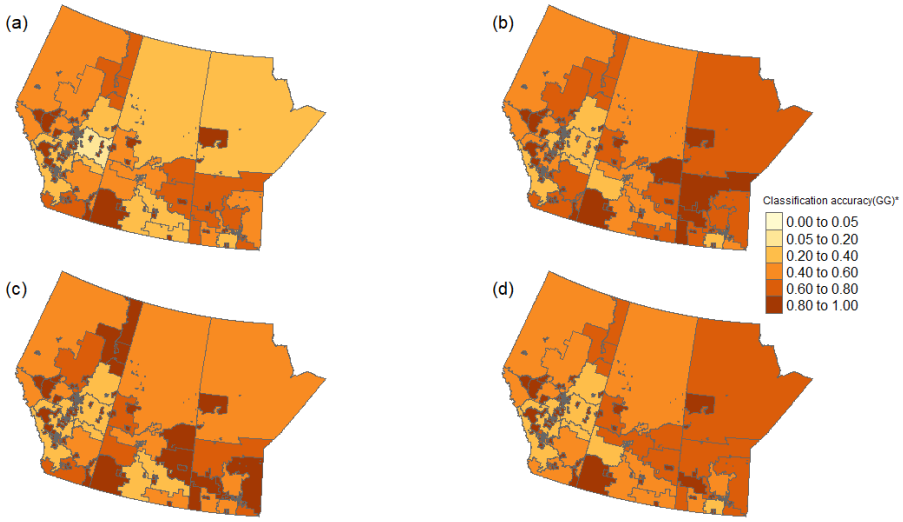

Figure 4.12 Classification accuracy of Getis-Ord Gi* (CA FSAs) (a) Detecting high-high clusters under the scenario of high-high clusters ; (b) Detecting high-high clusters under the scenario of the mixture of high-high and low-low clusters; (c) Detecting low-low clusters under the scenario of the mixture of high-high and low-low clusters; (d) Detecting high-high clusters under the scenario of the mixture of high-high clusters and high-low outliers.

Table 4.6 Mean Type I error for local Moran's I (LM) and Getis-Ord Gi* (GG*).

Data generation mechanism (DGM)	Regular grid		USA counties		CA FSAs	
	LM	GG*	LM	GG*	LM	GG*
D4: Null hypothesis (no spatial pattern)	0.00	0.01	0.00	0.01	0.01	0.01
D5: High-High clusters	0.01	0.03	0.01	0.02	0.02	0.04
D6: Mixture of high-high and low-low clusters	0.03	0.05	0.02	0.03	0.03	0.05
D7: Mixture of high-high clusters and high-low outliers	0.04	0.06	0.02	0.03	0.03	0.06

Table 4.7 Mean classification accuracy for local Moran's I.

Data generation mechanism (DGM)	Regular grid		USA counties		CA FSAs	
	Cluster accuracy	Outlier accuracy	Cluster accuracy	Outlier accuracy	Cluster accuracy	Outlier accuracy
D5: High-High clusters	0.73	-	0.92	-	0.71	-
D6: Mixture of high-high and low-low clusters	0.83	-	0.95		0.80	-
D7: Mixture of high-high clusters and high-low outliers	0.82	1.00	0.94	1.00	0.78	0.89

We calculate mean type I error (Table 4.6) and mean classification accuracy over all the units for local Moran's I (Table 4.7) and Getis-Ord Gi* (Table 4.8). The mean type I error is less than or equal to 0.05 under all DGMs except for the mean type I error (0.06) for Getis-Ord Gi* under the scenario of the mixture of high-high clusters and high-low outliers. Mean type I error for Getis-Ord Gi* is greater than or equal to the mean type I error of local Moran's I under each DGMs for all three areal structures. According to Table 4.7, the mean classification accuracy of detecting a cluster for local Moran's I is highest for USA counties and lowest for CA FSAs under all DGMs. The power of detecting clusters is higher under D6 and D7 compared to D5. The power of outlier detection for local Moran's I is 1 for both regular grid and USA counties. According to

Table 4.8, the power of detecting high-high and low-low clusters is higher for USA counties than for the grid or the CA FSAs. The power of detecting high-high clusters is higher under D6 and D7 compared to D5 which is similar to what we saw in local Moran's I. Classification accuracy in detecting clusters is higher for Getis-Ord Gi* than the local Moran's I for all the areal structures.

Table 4.8 Mean classification accuracy for Getis-Ord Gi*.

Data generation mechanism (DGM)	Regular grid		USA counties		CA FSAs	
	HH cluster accuracy	LL cluster accuracy	HH cluster accuracy	LL cluster accuracy	HH cluster accuracy	LL cluster accuracy
D5: High-High clusters	0.84	-	0.97	-	0.84	-
D6: Mixture of high-high and low-low clusters	0.94	0.94	0.99	0.99	0.91	0.90
D7: Mixture of high-high clusters and high-low outliers	0.93	-	0.98	-	0.89	-

CHAPTER 5: DISCUSSION

In this study, we simulate data under different data generation mechanisms to assess the performance of four spatial clustering methods: ANN, Ripley's K function, local Moran's I, and Getis-Ord Gi* statistics. All these methods are applied on three areal structures, a 20x20 regular grid, USA counties in 6 states, and CA FSAs in three provinces. As we know, ANN and Ripley's K functions are intended for point pattern data but in practice these functions are often used for areal data, violating a basic assumption about the distribution of the data under the null hypothesis. Our results show that the empirical type I error rates of ANN and Ripley's K are inflated for the simulated data regardless of the sample size. From the results, we see that the power of ANN to detect clusters for CA FSAs is higher than for regular grid and USA counties. One explanation of this is that many of the areal units in CA FSAs are very small and close together while some of the units are very large. As a result, the centroids of the smaller units are more likely to appear clustered relative to the centroids of the larger units. For the regular grid and the USA counties, the power of detecting clustering under DGM 2 (the single large cluster) is relatively low. Under this DGM, most of the observed units are clustered together, but the remaining units are further away from each other than would be expected under complete spatial randomness. It is possible that the contribution of the dispersed units to the ANN test statistic 'dilutes' the contribution from the clustered units and causes the loss of power. For the regular grid, we see that when sample size increases the power of detecting multiple clusters decreases rapidly. For the regular grid and the larger sample size, the expected

distance to the nearest neighbor is very close to the distance between centroids of adjacent points. Note that most of the units in regular grid has eight adjacent units and under multiple clustering, most points have an adjacent unit as their nearest neighbor. So, the ANN ratio is closer to 1. When the sample size is small, the expected distance to the nearest neighbor is noticeably larger than the distance between centroids of adjacent points and ANN ratio is less than 1. In this case, probability of rejecting the null hypothesis of CSR under multiple clusters is high, hence the power is high. We also see that the ANN gives different results depending on which window you use for the ANN calculation. The highly inflated type I error rate makes ANN an unreliable method for detecting spatial clustering in areal data.

Ripley's K function performs better in detecting clusters compared to ANN but at the same time, we see that the type I error is inflated. For the CA FSAs and USA counties, Ripley's K function rejects the null hypothesis of CSR in favor of clustering except for the smallest radius. For the regular grid, Ripley's K function rejects the null hypothesis in favor of dispersion. As the centroids of all the grid units are inherently dispersed, the observed units exhibit dispersion as well. We also see that the power of detecting clusters is unreliable for the USA counties and CA FSAs for the smallest radius. Even though we see that the power of Ripley's K is high for most of the scenarios, the type I error rates are also high. Due to the inflated type I error rate, the results from Ripley's K are not reliable when applied to areal data.

Unlike the ANN and Ripley's K, local Moran's I and Getis-Ord Gi* statistics can be applied to both point and areal data. Local Moran's I classifies observed units as part of a cluster or an outlier while Getis-Ord Gi* classifies units as part of either high or low

clusters. Both of these statistics are inherently linked to each other, and we typically expect similar results with both techniques in terms of which spatial units are significant. We see that the type I error rates are smaller for both statistics for all three areal structures. The type I error rate of local Moran's I is lower than that of Getis-Ord Gi*, though the type I error rate of both methods is near or below its nominal level. Overall, the power of detecting clusters is high for both of these statistics. However, Getis-Ord Gi* has a higher classification accuracy compared to local Moran's I. Overall, local Moran's I and Getis-Ord Gi* appear to be reliable for detecting spatial patterns in areal data when binary adjacency-based weights are used.

CHAPTER 6: CONCLUSION

In this study, we carried out a simulation study to assess the performance of four spatial clustering detection methods, the average nearest neighbor ratio, Ripley's K function, local Moran's I and Getis-Ord Gi* statistics. We applied these statistics to three different areal structures under different data generation mechanisms and calculated empirical type I error rate and power. Our findings suggest that the ANN and Ripley's K do not deliver reliable results when applied to areal data. In contrast, local Moran's I and Getis-Ord Gi* statistics appear to be reliable for detecting spatial patterns in areal data.

There are several areas for potential future work. We use binary adjacency-based weights to calculate local Moran's I and Getis-Ord Gi*. Future work could explore different weights (e.g. distance based) for these statistics and compare their performances. We could also simulate data from non-normal distributions, such as count or rate data. Different mechanisms for generating clustering and outliers could also be explored.

REFERENCES

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis, 27*(2), 93-115.

Baruch-Mordo, S., Breck, S. W., Wilson, K. R., & Theobald, D. M. (2008). Spatiotemporal distribution of black bear-human conflicts in Colorado, USA. *The Journal of Wildlife Management, 72*(8), 1853-1862.

Bone, C., Wulder, M. A., White, J. C., Robertson, C., & Nelson, T. A. (2013). A GIS-based risk rating of forest insect outbreaks using aerial overview surveys and the local Moran's I statistic. *Applied Geography, 40*, 161-170.

Clark, P. J., & Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology, 35*(4), 445-453.

Cressie, N. (1994). Models for spatial processes. *Statistical Methods for Physical Science, 28*(93), 124.

Dixon, P. M. (2014). R ipley's K function. *Wiley StatsRef: Statistics Reference Online*.

Getis, A., & Franklin, J. (1987). Second-Order Neighborhood Analysis of Mapped Point Patterns. *Ecology*, 473-477.

Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. In *Perspectives on spatial data analysis* (pp. 127-145): Springer.

Haase, P. (1995). Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction. *Journal of vegetation science, 6*(4), 575-582.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Jana, M., & Sar, N. (2016). Modeling of hotspot detection using cluster outlier analysis and Getis-Ord Gi* statistic of educational development in upper-primary level, India. *Modeling Earth Systems and Environment, 2*(2), 60.

Mansour, S. (2016). Spatial analysis of public health facilities in Riyadh Governorate, Saudi Arabia: a GIS-based study to assess geographic variations of service provision and accessibility. *Geo-spatial Information Science, 19*(1), 26-38.

Mollalo, A., Alimohammadi, A., Shirzadi, M. R., & Malek, M. R. (2015). Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. *Zoonoses and public health, 62*(1), 18-28.

Moore, D. A., & Carpenter, T. E. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic Reviews, 21*(2), 143-161.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability, 13*(2), 255-266.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(2), 172-192.

Ripley, B. D. (1979). Tests of'randomness' for spatial point patterns. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(3), 368-374.

Sayer, D., & Wienhold, M. (2013). A GIS-investigation of four early Anglo-Saxon Cemeteries: Ripley's K-function analysis of spatial groupings amongst graves. *Social Science Computer Review, 31*(1), 71-89.

Siebeneck, L. K., Medina, R. M., Yamada, I., & Hepner, G. F. (2009). Spatial and temporal analyses of terrorist incidents in Iraq, 2004–2006. *Studies in Conflict & Terrorism, 32*(7), 591-610.

Sokal, R. R., & Thomson, B. A. (2006). Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 129*(1), 121-131.

Songchitruksa, P., & Zeng, X. (2010). Getis–Ord spatial statistics to identify hot spots by using incident management data. *Transportation research record, 2165*(1), 42-51.

Whallon, R. (1974). Spatial analysis of occupation floors II: the application of nearest neighbor analysis. *American Antiquity, 39*(1), 16-34.

Zhang, C., Luo, L., Xu, W., & Ledwith, V. (2008). Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the total environment, 398*(1-3), 212-221.