Spatial Scan Statistics in Loglinear Models

Tonglin Zhang * and Ge Lin †

Purdue University and West Virginia University

Abstract

The spatial scan statistic based on likelihood ratio has been widely used in spatial disease surveillance and other spatial cluster detection applications. In order to better understand cluster mechanisms, we propose an equivalent model-based approach to the spatial scan statistic that unifies currently loosely coupled methods for including ecological covariates in the spatial scan test. In addition, we demonstrate the utility of the model-based approach with a Wald-based scan statistic that can account for overdispersion and heterogeneity in background rates. Our simulation and case studies show that both the likelihood ratio-based and Wald-based scan statistics are comparable with the original spatial scan statistic for cluster detection.

Key Words: bootstrap, heterogeneity, model-based scan statistic, overdispersion, spatial scan statistic, Wald statistic.

*Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066, Email: tlzhang@stat.purdue.edu

†Department of Geology and Geography, West Virginia University, Morgantown, WV 26506-6800, Email: glin@wvu.edu
1 Introduction

Modeling spatial phenomena and detecting spatial clusters are two traditions in spatial statistics. Spatial econometricians, regional scientists, and ecologists are mainly interested in the former for obtaining less biased substantively driven parameter estimates. Spatial epidemiologists and geographers are interested in the latter for detecting and quantifying localized clusters, particularly with regard to possible relationships between clustered phenomena and corresponding environments. The assessments of spatial patterns from both traditions often involve statistical tests for an overall spatial clustering as well as elevated local clusters. While these two traditions have been linked in spatial regressions for continuous or Gaussian data since Cliff and Ord’s book [2], they are rather separated when dealing with Poisson count data. Recently Lin and Zhang [11] attempted to bridge the two traditions for Poisson data based on Moran’s $I$ statistic that was originally designed for Gaussian data [14]. In this paper, we attempt to bridge the two traditions along this line based on Kulldorff’s spatial scan statistic that was originally designed for Poisson or binomial data [7].

We chose Kulldorff’s spatial scan statistics for two reasons. First, it is the most widely used cluster detection method for disease surveillance. It uses a moving circle of varying size to detect a set of clustered regions or points that unlikely happen by chance. Time and categorical covariates, such as age or sex, can be included as control variables. Since their publications, the two papers by Kulldorff and Nagarwalla in 1995 [6] and Kulldorff in 1997 [7] have garnered more than 150 and 200 citations respectively from the combined science and social science citation indices. Government, state and city agencies, such as the National Cancer Institute, Washington State Department of Public Health, New York City Department of Mental Health and Hygiene use SatSan, the software that performs the test for various retrospective and prospective cluster detection and surveillance. Regardless of retrospective and prospective cluster detection, there is a genuine need to incorporate
ecological covariates in the spatial scan statistic for better understanding cluster mechanisms. For instance, after identifying local breast cancer clusters by the spatial scan method, Roche et al. [18] compared local socioeconomic conditions within and outside of clustered areas. Recuenco et al. [17] followed a loosely coupled method [9], and examined spatial and temporal patterns of enzootic raccoon rabies with multiple covariates. They first fitted a regular Poisson regression with known ecological covariates, and then performed the spatial scan test by using the expected value from each spatial unit. In both applications, it is desirable to have a model-based spatial scan test that can either explain detected clusters with ecological covariates, or control for covariates.

Second, the spatial scan statistic is shown to be the most powerful for local cluster detection [8, 19], but “the probability of both rejecting the null hypothesis and detecting the true cluster correctly is a different matter” [20]. The Poisson distribution assumes that its variance equals its mean, but in reality they are often not equal due to missing variables ([1], P 7). Cases are overdispersed when their variability exceeding that predicted by the Poisson distribution. In the context of spatial phenomena, a common cause of overdispersion is spatial correlated and uncorrelated heterogeneities [4]. Many spatial test statistics including the spatial scan statistic are unable to treat overdispersion. On one hand, most of cluster detection methods are not model-based. Hence, many model-based treatments for overdispersion cannot be directly applied. On the other hand, when the underlying assumption is based on the Poisson distribution, a test statistic based on maximum likelihood is often difficult to compute in the presence of overdispersion or heterogeneity. Consequently, the scan statistic may have an unacceptable type I error probability in the presence of overdispersion [12]. In such a situation, the user has to decide if an alarm or signal of cluster is false or not. For example, dispersed cases were noticed in the syndromic surveillance of lower respiratory infection by the spatial scan in Eastern Massachusetts [5]. After the adjustment of known factors that caused overdispersion, the false alarm rate could be reduced as high as 30%. In addition to adjusting known factors, there are at least two ways to account for
overdispersion statistically. One is to add a parameter to a generalized linear model that allows an extra variation from the mean. The other way is to model spatial random effects. Both require a model-based approach that can incorporate an additional parameter into the spatial scan statistic.

There are many other situations where a model-based spatial scan statistic is desirable, and it is almost impossible to list all of them. For the purpose of the current study, we intend to develop a model-based approach to Kulldorff’s spatial scan statistic. In particular, we propose the model-based likelihood scan statistic $G^2_s$. In addition, we propose a new spatial scan statistic based on the Wald statistic to demonstrate the utility of the model-based approach. We show that the Wald-based spatial scan statistic does not suffer the complication of computing the likelihood function, and it can be used to detect local cluster in the presence of overdispersion. In the following section, we first briefly describe the spatial scan statistic and then propose the model-based scan statistics: the likelihood ratio (deviance) scan and the Wald scan statistics. We show that in the absence of covariates, the likelihood ratio scan statistic is equivalent to Kulldorff’s spatial scan statistic. In addition, the model-based scan statistic can be effective when considering overdispersion. In section 3, we present simulation and case studies by using infant mortality data from Guangxi Province in China. We first compare the powers among the two model-based tests via Monte Carlo simulations, and then present the study of infant mortality. Finally, in section 4, we provide some concluding remarks.

2 From the spatial scan to model-based scan statistics

*Kulldorff scan test.* Kulldorff developed the spatial scan statistic by combing time series scan statistic[15] and spatial search machine methods[16]. The spatial scan test uses a moving circle of varying size to find a set of regions or points that maximizes the likelihood ratio test for the null hypothesis of a purely random Poisson or Bernoulli process. Time can be included for a space-time
scan test. Categorical covariates, such as age group and sex, can be included as control variables. Conceptually it is straightforward to extend the circular scan window to other shapes, but only until recently the elliptical shape was added to SatScan, the software that hosts the family of spatial scan test functions.

The spatial scan statistic requires that 1) we have a set of \( m \) locations or subregions in a study area, and 2) each location \( i \) has a number of case counts \( y_i \) and corresponding at-risk population \( n_i \). Assume that the counts are independent Poisson random variables with expected value \( n_i \theta_i \), where \( \theta_i \) is the unit specific relative risk. The scan statistic is to detect any spatial regions within a cluster \( C \) in which counts are significantly higher than expected. The test compares the total number of disease counts, \( y_c = \sum_{i \in C} y_i \), within \( C \) with the total number of counts, \( y_\bar{c} = \sum_{i \notin C} y_i \), outside of \( C \), given the corresponding at risk populations inside and outside of \( C \), denoted by \( n_c = \sum_{i \in C} n_i \) and \( n_\bar{c} = \sum_{i \notin C} n_i \) respectively. Assuming that \( \theta_i = \theta_c \) for \( i \in C \) and \( \theta_i = \theta_0 \) for \( i \notin C \), and contrasting the null hypothesis of \( H_0 : \theta_c = \theta_0 \) against the alternative hypothesis of \( H_1 : \theta_c > \theta_0 \), Kulldorff developed the likelihood ratio as:

\[
\Lambda_C = \frac{\max_{\theta_c > \theta_0} L(C, \theta_c, \theta_0)}{\max_{\theta_c = \theta_0} L(C, \theta_c, \theta_0)} = \left( \frac{y_c/n_c}{y/n} \right)^{y_c} \left( \frac{y_\bar{c}/n_\bar{c}}{y/n} \right)^{y_\bar{c}}, \tag{1}
\]

when \( y_c/n_c \geq y_\bar{c}/n_\bar{c} \) and \( \Lambda_C = 1 \) otherwise, where \( L(C, \theta_c, \theta_0) \) is the likelihood function. Since \( C \) is unknown, it can be treated as a parameter and the maximum likelihood ratio test statistic for unspecified spatial cluster \( C \) is given by

\[
\Lambda = \max_{C \in \mathcal{C}} \Lambda_C \tag{2}
\]

where \( \mathcal{C} \) is the collection of candidates of possible spatial clusters. Since the null distribution of \( \Lambda \) is not analytically tractable, Kulldorff recommends to compute the \( p \)-value by Monte Carlo simulations.

Like most test statistics, the spatial scan statistic also has some limitations. For instance, the spatial scan statistic cannot directly include ecological covariates, but Kulldorff recommends
a loosely coupled method that uses the expected values from a loglinear model that includes ecological covariates [9]. In addition, it does not take overdispersion into consideration. Although it is rather straightforward to extend the spatial scan statistic to an equivalent model-based spatial scan statistic, such an effort will likely save a lot of troubles for those interested in including ecological covariates. Here we propose a model-based spatial scan test based on a loglinear model. We show that the model-based likelihood ratio scan statistic is essentially equivalent to the spatial scan statistic in the absence of ecological covariates.

It is, however, more complicated to find a way to deal with overdispersion in the model-based scan approach. A common way to account for overdispersion is to extend the Poisson regression to the negative binomial regression. However, this method usually implies a specific mechanism for which the relationship between the mean and variance functions is known. For spatially autocorrelated data, overdispersion is often caused by unknown factors, the negative binominal distribution cannot be extended to allow spatial correlated structure [21]. Loh and Zhu offered an MCMC method that adjusts the calculation of \( p \)-value in the spatial scan test for spatial autocorrelation [12]. However, since this method was mainly concerned with spatial correlation, it only addressed the part of heterogeneity problem without getting into model correction. In a more general case, an overdispersion parameter that allows the variance greater than the mean can be introduced in a Poisson regression [3], so that some existing estimation method for overdispersion, such as the Quasi-Likelihood method [13], can be extended to the spatial scan test. Since the Wald-statistic is a computationally tractable and proven to be effective for accounting for overdispersion, we propose two Wald-based spatial scan tests. One can be used similarly as the spatial scan statistic. The other can be used when considering overdispersion.

*Model-based spatial scan statistic.* We use the concept of relative risk to introduce a spatial
loglinear model:

\[
\log(\theta_i) = \mathbf{x}_i^T \beta + \alpha_i
\]  

where \( \beta \) is the vector of coefficients of parameters, \( \alpha_i \) is a cluster indicator defined as \( \alpha_i = \alpha_c \) if \( i \in C \) and \( \alpha_i = \alpha_0 \) if \( i \notin C \). By the constraint, we can set \( \alpha_0 = 0. \) Under the null hypothesis of \( H_0 : \alpha_c = 0, \) model (3) becomes

\[
\log(\theta_i) = \mathbf{x}_i^T \beta.
\]  

(4)

Under the alternative hypothesis of \( H_1 : \alpha_c > 0 \) for some \( C, \) model (3) becomes

\[
\log(\theta_i) = \mathbf{x}_i^T \beta + \alpha_c I_{i \in C}.
\]  

(5)

When \( C \) is given, the MLE (maximum likelihood estimate) of \( \beta \) under the null hypothesis and the MLE of \( \alpha_c \) and \( \beta \) under the alternative hypothesis can easily be derived (see [1, 3]). Let \( \hat{\beta}_0 \) be the MLE under the null hypothesis and \( \hat{\alpha}_c \) and \( \hat{\beta}_1 \) be the MLE under the alternative hypothesis. The predicted counts based on \( H_0 \) and \( H_1 \) can be derived by \( \hat{y}_i^0 = n_i e^{\mathbf{x}_i^T \hat{\beta}_0} \) and \( \hat{y}_i^1 = n_i e^{\mathbf{x}_i^T \hat{\beta} + \hat{\alpha}_c I_{i \in C}} \) respectively. The \(-2\) times the logarithm of the likelihood ratio statistic (the deviance statistic) conditional on cluster \( C \) is

\[
G^2_C = 2 \sum_{i=1}^{m} y_i \log \frac{\hat{y}_i^1}{\hat{y}_i^0}.
\]  

(6)

If \( \mathbf{x}_i^T \beta \) only includes the intercept term \( \beta_0, \) then under \( H_0, \) we have \( \log(\theta_0) = \beta_0, \) and under \( H_1, \) we have \( \log(\theta_i) = \beta_0 + \alpha_c I_{i \in C}. \) We now show that \( G^2_C \) given by equation (6) is equivalent to the Kulldorff likelihood ratio function \( \Lambda_C \) given by equation (1). The maximum likelihood estimates are: \( \hat{\beta}_0^0 = \log(y/n) \) under \( H_0, \) and \( \hat{\beta}_0^1 = \log(y_e/n_e) \) and \( \hat{\alpha}_c = \log(y_c/n_c) - \log(y_e/n_e) \) if \( y_c/n_c \geq y_e/n_e, \) or \( \hat{\beta}_0^1 = \log(y/n) \) and \( \hat{\alpha}_c = 0 \) if \( y_e/n_c < y_e/n_e \) under \( H_1. \) Then \( \hat{y}_i^0 = y n_i/n \) and

\[
\hat{y}_i^1 = \begin{cases} 
\frac{y m_i}{n} I_{i \in C} + \frac{y_e}{n_e} I_{i \notin C}, & \text{when } y_e/n_c \geq y_e/n_e \\
\frac{y m_i}{n}, & \text{when } y_e/n_c < y_e/n_e.
\end{cases}
\]
Therefore, we have

\[ G^2_C = 2 \sum_{i \in C} y_i \log \left( \frac{y_i / n_c}{y / n} \right) + 2 \sum_{i \notin C} y_i \log \left( \frac{y_i / n_c}{y / n} \right) = 2 y_c \log \left( \frac{y_c / n_c}{y / n} \right) + 2 y_{\bar{c}} \log \left( \frac{y_{\bar{c}} / n_{\bar{c}}}{y / n} \right), \]

if \( y_c / n_c \geq y_{\bar{c}} / n_{\bar{c}} \) and \( G^2_C = 0 \) otherwise, which implies \( G^2_C = 2 \log \Lambda_C \).

Likewise, we define the Wald statistic as

\[ Z_C = \frac{\hat{\alpha}_c}{\hat{\sigma}_{\hat{\alpha}_c}}, \tag{7} \]

where \( \hat{\sigma}_{\hat{\alpha}_c} \) is the estimate of the standard error of \( \hat{\alpha}_c \) under the alternative hypothesis.

Following Kulldorff, we define the model-based deviance scan statistic as

\[ G^2_s = \max_{C \in C} G^2_C \tag{8} \]

and the Model-based Wald scan statistic as

\[ Z_s = \max_{C \in C} Z_C. \tag{9} \]

It should be pointed out that the likelihood-based spatial scan statistic \( G^2_s \) is analytically more appealing, but the Wald-based spatial scan statistic \( Z_s \) is more flexible. Since \( Z_s \) only involves the estimates of the cluster parameter \( \alpha_c \) and its standard error, and most model fitting procedures for GLMMs can provide parameter estimates and their standard errors, \( Z_s \) can be easily modified for many spatial GLMMs. If, for instance, a Poisson regression is approximated by a negative binomial regression, then \( Z_s \) can be modified by taking \( \hat{\alpha}_c \) and \( \hat{\sigma}_{\hat{\alpha}_c} \) as the estimated value from the negative binomial regression model. If a spatial correlated or uncorrelated random effect is implemented in the Poisson regression model, then \( Z_s \) can be modified under spatial GLMMs.

In both cases, \( Z_C \) can be derived by calculating the corresponding parameter estimate and its standard error based on equation (7). Due to its flexibility, the proposed \( Z_s \) is a nice addition to the spatial scan statistic family. In the following, we only provide a simple modification to demonstrate its utility for overdispersion.
Modification for overdispersion. Overdispersion can occur in spatial Poisson process either because spatially correlated data, or because spatially varied cases and populations from a large number of spatial units. To gauge and to reduce the potential effect of overdispersion, a dispersion parameter denoted by $\phi$ can be introduced similar to the general overdispersion parameter in a Poisson regression. When $\phi = 1$, it is a regular Poisson regression without overdispersion. When $\phi > 1$, it is a Poisson regression with overdispersion. If we know the extent of spatially correlated data, an adjustment of deviance via $\phi$ for a $p$-value can be effective [12]. When we know the exact relationship between the variance and the expected value ([1], p 559-565), a negative binomial regression can also be effective. In the absence of these information, an estimate of the dispersion parameter can be used [13]:

$$\hat{\phi} = \max(\frac{X^2}{df_{res}}, 1),$$

(10)

where $X^2$ is the Pearson $\chi^2$ statistic and $df_{res}$ is the residual degrees of freedom. When an overdispersion parameter is considered, the estimated variance of $\hat{\alpha}_c$ is adjusted by $\hat{\phi}^{1/2}\hat{\sigma}_{\alpha_c}$ and the Wald statistic testing for the significance of cluster $C$ conditional on cluster $C$ is modified as

$$Z_{C,o} = \frac{\hat{\alpha}_c}{\hat{\phi}^{1/2}\hat{\sigma}_{\alpha_c}}.$$  

(11)

The model-based Wald scan statistic with adjustment of overdispersion parameter is then defined as

$$Z_{s,o} = \max_{C \in C} Z_{C,o}.$$  

(12)

Note that under the loglinear model framework, the above unadjusted $Z_s$ does not consider the overdispersion parameter $\phi$ and thus it always assigns the dispersion parameter equal to 1. The $Z_{s,o}$ statistic, which is treated as a modification of $Z_s$, considers the overdispersion parameter.

Similar to Kulldorff’s spatial scan statistic $\Lambda$, the null distributions of $G^2_s$, $Z_s$ and $Z_{s,o}$ are not analytically tractable. We, therefore, follow Kulldorff to compute their $p$-values by the bootstrap
method. The null hypothesis of no clustering is rejected for large $G^2_{s}$, $Z_{s}$ or $Z_{s,o}$ value. The corresponding bootstrap $p$-value is the rank of the real observed values in the corresponding bootstrap distribution. In the same way as in Kulldorff [7], we calculate the $p$-value conditionally on the total number of counts by the following algorithm.

**Bootstrap for the $p$-values of $G^2_{s}$, $Z_{s}$ and $Z_{s,o}$:**

i) Calculate the observed values of $G^2_{s}$, $Z_{s}$ and $Z_{s,o}$ (denoted by $G^2_{s,0}$, $Z_{s,0}$, and $Z_{s,o,0}$ respectively) according to equations (8), (9) and (12) respectively.

ii) Derive the predicted counts $\hat{\mu}_i$ under the null model (4).

iii) Generate $K$ independent multinomial random variables with total counts equal to $Y$ and proportional vector equal to $(\hat{\mu}_1/Y, \cdots, \hat{\mu}_m/Y)$. Calculate the simulated values of $G^2_{s}$, $Z_{s}$ and $Z_{s,o}$ (denoted by $G^2_{s,k}$, $Z_{s,k}$ and $Z_{s,o,k}$ respectively for $k = 1, \cdots, K$) for each generation.

iv) Report the $p$-values of $G^2_{s}$, $Z_{s}$ and $Z_{s,o}$ by #$\{G^2_{s,k} \geq G^2_{s,0} : k = 0, \cdots, K\}/(K + 1)$, #$\{Z_{s,k} \geq Z_{s,0} : k = 0, \cdots, K\}/(K + 1)$ and #$\{Z_{s,k,o} \geq Z_{s,o,0} : k = 0, \cdots, K\}/(K + 1)$ respectively, where #$A$ represents the number of elements in set $A$.

Both model-based $G^2_{s}$ and $Z_{s}$ statistics have their roles in spatial scan for clusters. $G^2_{s}$ is conceptually equivalent to the loosely coupled spatial scan statistic that can include ecological covariates. In the presence of overdispersion, however, it is rather complicated to incorporate $\hat{\phi}$ to $G^2_{s}$. In this case, the model-based Wald scan statistics ($Z_{s}$ and $Z_{s,o}$) can be effective.

3 Simulation and case study

In both simulation and case studies, we used real world data based on 110 counties in Guangxi, China. Guangxi is one of five autonomous regions for ethnic minorities in China, and is the home of Zhuang ethnic group. We obtained the county level infant birth and death data from the 2000
Chinese Census. Infant mortality rates vary substantially among these counties, ranging from 248 to 7260 per 100,000 (Figure 1). Guangxi for the most part is mountainous and is considered a less developed region in southwestern China. The county level elevations range from 20 to 1,140 meters above the sea level, lower in the southeast and higher toward the west. In the absence of other variables, elevation is a good measure of access to care, education and other socioeconomic resources. For this reason, we included elevation in both simulation and the case study.

3.1 Simulation

We evaluate the type I error rates and the power functions of the proposed statistics. For a pure spatial situation, the model-based likelihood ratio scan statistic is equivalent to the original spatial scan statistic, and its statistical powers have been intensively evaluated [8]. For this reason, we opted to evaluate the performance of these model-based scan statistics by including an ecological covariate. For simplicity, we used 110 counties to generate the spatial weight matrix based on spatial adjacency, and used infant births \( n_i \) and deaths \( Y_i \), and elevations \( x \) to fits a baseline model. Assume that \( E(Y_i) = \theta_i n_i \), we fitted a baseline model with elevation and its quadratic term as ecological covariates and the fitted model is:

\[
\log(\theta_i) = -4.160 + 2.07 \times 10^{-3} x - 1.27 \times 10^{-6} x^2.
\]

The coefficient of the quadratic term in the fitted model was negative. The maximum predicted mortality rate was \( (2.07 \times 10^{-3})/(2 \times 1.27 \times 10^{-6}) = 815 \). Since the elevations of most counties are lower than 815 meters (103 out of 110), model (13) suggests that the infant mortality rate increased as elevation increased.

Model (13) serves as a baseline model for generating a new response and a local cluster term \( (w_{54}) \). We inserted a seven-county local cluster \( C \) in the baseline model with its center at a
non-border county Pingnan, and the new response variable $Y_i'$ was then generated from

$$
\log(\theta_i') = -4.160 + 2.07 \times 10^{-3} x - 1.27 \times 10^{-6} x^2 + \delta I_{i \in C},
$$

where $\theta_i' = E(Y_i')/n_i$, $I_{i \in C}$ is the indicator function, which is 1 if $i \in C$ and is 0 otherwise. We used $\delta$ to measure the strength of the cluster from 0 to 20 or 20% more than the relative risk outside of the cluster. Since the model was exactly Poisson, there was no overdispersion for the simulated data.

In the simulation, we compared the type I error rates $G^2_s, Z_s$ and $Z_{s,o}$ when $\delta = 0$ and their power functions as $\delta$ increased (see Table 1). We generated random Poisson counts and calculated $p$-values for $G^2_s, Z_s$ and $Z_{s,o}$ based on 999 random replications of the null distribution from Monte Carlo hypothesis testing. We repeated the procedure 1,000 time for each selected $\delta$ value. Since a spatial cluster can be treated as a spatial association term or an explanatory variable in the loglinear model, we scanned all counties and searched for the cluster that yielded

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$G^2_s$</th>
<th>$Z_s$</th>
<th>$Z_{s,o}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>6.0 0.0</td>
<td>5.8 0.0</td>
<td>4.3 0.0</td>
</tr>
<tr>
<td>0.025</td>
<td>7.1 8.5</td>
<td>6.8 10.3</td>
<td>6.8 14.7</td>
</tr>
<tr>
<td>0.05</td>
<td>14.9 39.6</td>
<td>14.9 38.9</td>
<td>12.6 42.1</td>
</tr>
<tr>
<td>0.075</td>
<td>37.4 65.2</td>
<td>37.9 66.2</td>
<td>39.4 63.5</td>
</tr>
<tr>
<td>0.1</td>
<td>71.6 82.4</td>
<td>72.3 82.2</td>
<td>73.5 81.9</td>
</tr>
<tr>
<td>0.125</td>
<td>93.9 90.2</td>
<td>94.3 90.7</td>
<td>95.0 90.3</td>
</tr>
<tr>
<td>0.15</td>
<td>99.1 96.5</td>
<td>99.2 96.3</td>
<td>99.3 96.7</td>
</tr>
<tr>
<td>0.175</td>
<td>100.0 99.3</td>
<td>100.0 99.3</td>
<td>100.0 99.3</td>
</tr>
</tbody>
</table>
the largest likelihood ratio or Wald z-value. We fixed the significance level at 0.05. We evaluated the detectability of inserted cluster by the percentage of p-values that were less than or equal to the significance level, or the rejected rates of the null hypothesis of no spatial clustering. In addition, we assessed the specificity of detected local cluster by the percentage of correctly locating the cluster center among the total significant p-values. The results are shown in table (1), where “D” represents detectability and “w54” represents location specificity in percentage.

We found that all three model-based scan statistics had an acceptable type I error probabilities, with 6.0%, 5.8% and 4.3% respectively for $G^2_s$, $Z_s$ and $Z_{s,o}$. As $\delta$ increased, the power functions of detectability increased rapidly. When the relative risk inside the cluster was 0.175 more than the outside the cluster, the powers for all three were 100%. In addition, when the relative risk was 0.15 greater, there were more than 95% chances that the correct cluster center were identified.

To briefly summarize, the Wald-based spatial scan statistics performed well in comparison to the likelihood ratio based scan statistic. All three had acceptable type I error probabilities and power functions when adjusting for ecological covariates. When there was no overdispersion, the power functions of the three model-based spatial scan statistics had almost identical results as the strength of the cluster increased. They were all able to detect the existence of the cluster with a moderate cluster strength, and they were able to identify the location of the detected cluster 99% times when the relative risk within the cluster was 1.175 times of the outside relative risk. Even though we did not evaluate the pure spatial process without any ecological covariates, we expect the same results based on the formulations for $G^2_s$ and $Z_s$.

### 3.2 Case Study

As mentioned earlier, Guangxi is a provincial level autonomous region for ethnic minority Zhuang. Higher infant mortality in the region is a sensitive political issue for the central government that...
tries to improve the living standards for all ethnic groups and nationalities. It is already known
that access to adequate primary care is an issue and the availability of primary care is greater in
lower land counties than in highland and mountainous counties. An interesting question is that
in addition to access to care, which can be approximate by the lack of access due to elevation,
what other factors might play a role for high infant mortality in the region. One of coauthors was
from the region, and our task was to first identify local pocket of counties by controlling for the
elevation effect, and then turn our findings to provincial health officials for further health disparity
analyses.

We first fitted infant mortality rates with a loglinear model that includes a quadratic elevation
term as:

$$\log(\theta_i) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

where $x$ represented the county level elevation in meters. The estimated values were $\hat{\beta}_0 = -4.16$,
$\beta_1 = 2.07 \times 10^{-3}$ and $\hat{\beta}_2 = -1.27 \times 10^{-6}$, with standard errors $2.22 \times 10^{-2}$, $1.21 \times 10^{-4}$ and
$1.23 \times 10^{-7}$. Since both the linear term and quadratic term of elevation were highly significant,
we considered the following model for cluster detection:

\[ \log(\theta_0) = \beta_0 + \beta_1 x + \beta_2 x^2 + \alpha I_{i \in C}, \]

where \( C \in C \) and \( C \) is the collection of all candidates of circular local cluster.

In the analysis, we also considered potential overdispersion by using both \( Z_{s,o} \) and model-based Pearson residual Moran’s \( I \) or \( I_{PR} \) [11]. The original Moran’s \( I \) is for a continuous variable, and it does not suffer overdispersion. Since \( I_{PR} \) is based on the same formulation, it is not sensitive to overdispersion either. Based on this property and the fact that \( I_{PR} \) can be derived from the same loglinear model used for \( G^2_s \) and \( Z_s \), we can use \( I_{PR} \) to check for the consistency of \( Z_{s,o} \).

We used a stepwise scan method searching for local clusters by treating a local cluster as an explanatory variable. This method is identical to the spatial scan method for detecting the first cluster. However, if multiple clusters exist, our method to scan for the secondary cluster might be slightly different from the spatial scan statistic. In SatScan, the secondary cluster is searched by considering the first cluster. In our method, when the first cluster is identified, its effect is taken out by the first cluster spatial association term, and an additional cluster term is used to scan and test for the existence of an additional cluster. The stepwise scan stops if the test statistic is no longer significant [11]. In both SatScan and our stepwise search methods, there are no multiple testing problem.

In the scan process, the first cluster was signaled with a \( p \)-value less than 0.001 for \( G^2_s, Z_s, Z_{s,o} \) and \( I_{PR} \) (see Table 2). This cluster is close to the border with Vietnam, and the area had dilapidated infrastructure for years due to the 1979 Sino-Vietnam Border War nearby. The second cluster was signaled with a \( p \)-value less than 0.001 for \( G^2_s \) and \( Z_s \). The \( p \)-values for \( I_{PR} \) and \( Z_{s,o} \) were much weaker in the range between 0.01 and 0.05. Even though the second cluster was not far from the Vietnam border either, most counties were not near the war zone. However, due to the war and prewar refugees who flooded into China from Vietnam, the counties around were a
government designated settlement area with a large number of influx of ethnic Chinese refugees from Vietnam. It was likely that after 20 years, some refugees resettled elsewhere, but many of them, especially the poor still lived in government sponsored farms.

The signals for the potential third cluster were mixed. Even though the third cluster was signaled from $G^2_s$ and $Z_s$ scan statistics, the results from $Z_{s,o}$ and $I_{PR}$ were not significant. The overdispersion parameters were greater than 20 among the first three models, signaling a strong effect. Since only $Z_{s,o}$ considers the adjustment of the overdispersion problem, and the result is consistent with that from $I_{PR}$, the third possible cluster was much less trustworthy. Based on $Z_{s,o}$, the stepwise spatial scan procedure stopped concluding the existence of two clusters centered at Fangcheng Qu and Pubei Xian (see Figure 1). The final model, therefore, is:

$$\log(\theta_i) = -4.52 + 3.24 \times 10^{-3} x - 2.13 \times 10^{-6} + 0.927 w_{45} + 0.365 w_{51}.$$ 

Based on the final model with two spatial association terms $w_{45}$ and $w_{51}$, we calculated parameter estimates for elevation and its quadratic terms, $3.24 \times 10^{-3}$ and $-2.13 \times 10^{-6}$ respectively (with corresponding $z$-values 24.31 and $-16.31$ respectively). With the parameter estimate of overdispersion $\hat{\phi} = 20.16$, the adjusted $z$-values were 5.42 and $-3.65$, which were highly significant (with $p$-values less than 0.001). Based on these parameter estimates, we further examined mortality variation inside and outside of the clusters. The odds ratio of the first and second clusters were
2.53 and 1.44, respectively. Taking elevation into consideration the mortality rates of the counties within the first and second cluster were 153% or 44% higher than the expected value. Since the elevation that corresponds to the maximum predicted mortality rate was about 760 meters, and there were only 12 counties above this level, the two detected clusters were the net effect of a general positive relationship between elevation and infant mortality.

4 Discussion and Concluding Remarks

Spatial analysts are often interested in cluster detection in factors that contribute to the detected clusters. Previously, researchers have resorted the loosely coupled method that first models ecological covariates, and then used the expected values to perform spatial scan statistic for cluster detection. In this article, we have directly extended the spatial scan statistic of Kulldorff to include ecological covariates, and integrated the loosely coupled steps in the spatial scan statistic into a single model-based statistic, $G_s^2$. In the absence of ecological covariates, $G_s^2$ is equivalent to the spatial scan statistic. Both simulation and case studies show that the proposed $G_s^2$ statistic has desirable properties with an acceptable type I error probabilities and power functions. Similar to disease mapping, the method can be used to display residual information while controlling for other effects, such as overdispersion. Unlike disease mapping, a spatial explicit cluster term can be parameterized for its shape and size.

However, since $G_s^2$ is difficult to be used to account for overdispersion that could cause false cluster signals, we also proposed model-based scan statistics based on the Wald tests–$Z_s$, $Z_{s,o}$. Our initial evaluations suggest that $Z_s$ can be used in the same way as $G_s^2$, while $Z_{s,o}$ is an standard way to account for overdispersion. Our case study of infant mortality showed that a third cluster could be signaled if overdispersion were not considered, and the result from $Z_{s,o}$ signaled that the third one could be a false alarm. In this regard, the proposed Wald statistic $Z_{s,o}$ can complement
the spatial scan statistic. Since the Wald-based spatial scan statistic does not need to compute the marginal likelihood function in GLMMs, and there are many approximate computing methods (e.g. quasi-likelihood, the GEE method [10]), the Wald-based spatial scan statistic can be implemented in many circumstances for spatially clustered data. It should be noted that even though the Wald statistic used in $Z_{s,o}$ is widely recognized in practice, it does not involve in any spatial operations. If we can identify spatial overdispersion, a spatial explicit or implicit measure can be developed. Future work should explore various spatial explicit overdispersion measures and adjustment methods. In this regard, our proposed $Z_{s,o}$ provides an approach for further extending Wald-based statistics to a wide range of models and model-based test statistics.

The model-based spatial scan statistics provide more flexibilities than the spatial scan statistic. Recently, considerable efforts have been spent on refining different shapes for the spatial scan tests [20]. The proposed model-based method can treat the detection of cluster with different shapes as a modeling process. For instance, we identified two clusters in southern Guangxi that are adjacent to each other which can be covered roughly by an elliptical shape. This information can then feed the model to change the scan from the perfect circle to an ellipse. Even though we demonstrate our method and test statistics based on Kulldorff’s spatial scan test, the idea can be used in other testing method for spatial clusters for count data. As the Wald $z$-value of an unknown parameter can be easily derived under a variety of statistical models, the modifications of Wald-based spatial statistic for local cluster can be applied to modeling spatial clusters and clustering in many generalized linear mixed effect models. Since the spatial scan statistic has been extensively used in many fields [9], the flexibility gained from the likelihood ratio-based spatial statistic to Wald-based spatial statistics will enable spatial analysts to account for ecological covariates, overdispersion and spatial random effects. By linking cluster detection to cluster mechanism discovery, we hope that the model-based approach will find wide applications for Poisson count data.
References


ing clusters. *International Journal of Health Geographics*, available on “http://www.ij-
healthgeographics.com/content/pdf/1476-072X-4-11.pdf”.

tial variation in risk. In *Spatial Epidemiology: Methods and Applications*, Elliott P., Wakefield,