Spatial Clusters in a Global-Dependence Model

Tai-Chi Wang[a,*], Ching-Syang Jack Yue[a]

[a]*Department of Statistics, National Chengchi University. NO.64, Sec.2, ZhiNan Rd.,Wenshan District,Taipei City 11605, Taipei, Taiwan, R.O.C*

Abstract

Spatial data often possess multiple components, such as local clusters and global clustering, and these effects are not easy to be separated. In this study, we propose an approach to deal with the cases where both global clustering and local clusters exist simultaneously. The proposed method is a two-stage approach, estimating the autocorrelation by an EM algorithm and detecting the clusters by a generalized least square method. It reduces the influence of global dependence on detecting local clusters and has lower false alarms. Simulations and the sudden infant disease syndrome data of North Carolina are used to illustrate the difference between the proposed method and the spatial scan statistic.

Keywords: Local Cluster, Spatial Global Dependence, Conditional Autocorrelated Regressive Model, Spatial Scan Statistic, EM Estimates, Generalized Least Square

1. Introduction

In spatial data analysis, one of the frequently discussed issues is the relationship between geographical locations, that is, the identification of spatial patterns. The particular interest is whether certain locations are significantly different from other locations in the aspect of statistical testings. Besag and Newell (1991) categorized such tests into two types: general tests and focused tests. For the general tests, it can be further categorized as global clustering

[*] Corresponding author. Address: Department of Statistics, National Chengchi University. NO.64, Sec.2, ZhiNan Rd.,Wenshan District,Taipei City 11605, Taipei, Taiwan, R.O.C. Tel.:+886-2-29393091.ext81144; fax.:+886-2-8237-0079

Email address: taichi@alumni.nccu.edu.tw (Tai-Chi Wang)

and cluster detection tests. Kulldorff et al. (2006) gave a great amount of references of these tests.

Global clustering tests, such as Moran's I statistic and Geary's C statistic, are concerned with global clustering patterns. Global clustering patterns can be modeled by using spatial autoregressive models or conditional autoregressive models (Besag, 1974; Cressie, 1993). On the other hand, cluster detection tests are used to determine if some attributes of one or more subregions, such as incidence rates of disease, are unusually large, that is, to identify hot spots. Getis and Ord (1992) and Anselin (1995) discussed several statistics to test local dependence. In recent years, spatial cluster detection methods have been widely applied to many different fields.

However, the efficiency of cluster detection methods is often data-dependent. Among these methods, the spatial scan statistic (SaTScan) (Kulldorff and Nagarwalla, 1995) is perhaps the most popular and is considered to be quite effective in many instances. For example, Huang et al. (2008) had compared several cluster detection methods, such as circular and elliptic spatial scan statistics (SaTScan), flexibly shaped spatial scan statistics, Turnbull's cluster evaluation permutation procedure, local indicators of spatial association, and upper-level set scan statistics. They found that the SaTScan had the best performance in several synthetic cluster patterns. Although the past studies have shown that the SaTScan is quite effective in detecting spatial clusters (Kulldorff et al., 2003; Takahashi and Tango, 2006; Huang et al., 2008), few of these studies (and other cluster detection methods) discuss the performance of cluster detection in case of global spatial autocorrelation. Besides, it should be noted that the SaTScan relies on the Monte Carlo method to decide the significance of clusters, and the global spatial autocorrelation can distort the Monte Carlo results. Intuitively, cluster detection can be expected to be more accurate when considering the global dependence.

It should be noted that using different cluster analysis methods, such as globally autoregressive models and cluster detection methods, can also produce different interpretations. For example, Cressie and Read (1989) discussed spatial autoregressive models on the sudden infant death syndrome (SIDS) data from 1974–1984 for the counties of North Carolina and concluded that the errors show some (nonsignificant) spatial dependent structure. On the other hand, Kulldorff (1997) identified several local clusters using the SaTScan for the same data but with different combinations of all years. In other words, a pattern of one geographical scale can be identified as another spatial pattern.

In identifying spatial cluster patterns, almost all cluster detection methods assume that the data are independent rather than dependent. However, the result of cluster detection may probably be affected by the local effects, as well as global dependence. Ord and Getis (1995) showed that "when global autocorrelation exists, local pockets are harder to detect." Ord and Getis (2001) said, "If existing tests are applied without regard to global autocorrelation structure, type I errors may abound." They provided the local O statistic, which can accommodate spatial parameters identified from variograms and correlograms, to detect local clusters. However, their method did not consider that local clusters also affect the estimate of autocorrelation. Kulldorff (2006) depicted the difficulties of identifying these two patterns and gave a general framework for testing the spatial randomness. Lawson (2006) differentiated these two effects and gave more specific definitions of them. Although there are many articles describing the possibility of existing multiple patterns in spatial data, the solution to the model involved local clusters and global dependence is rarely mentioned. In this study, in addition to showing the difficulty of disentangling these two effects, we propose a method including both local clusters and global clustering.

In this paper, we propose a cluster detection approach that deals with global spatial dependence. Before introducing the proposed method, we will first review the conditional autoregressive spatial model (or the conditionally specified spatial Gaussian model) and the SaTScan in Section 2. In addition, we will evaluate the performance of the SaTScan in case of spatial autocorrelation. Then, the proposed approach is introduced, together with the EM estimates, the scanning procedure, and the Monte Carlo testing for handling both global dependence and clusters in Section 3. Simulations and an empirical study (the SIDS data of North Carolina) are used to evaluate the proposed methods in Sections 4 and 5. We will present comments and discussions on the proposed approach in the final section.

2. Spatial Models and the SaTScan

We first introduce the concepts of global dependence and cluster detection. Regarding to global dependence, we will provide a brief introduction of spatial models. For a complete introduction to spatial autoregressive models, please refer to Chapter 6 of Cressie (1993). To identify spatial clusters, the concept of the SaTScan is briefly introduced, and a detailed discussion of which can be found in Kulldorff's work (Kulldorff, 1997). Furthermore, we will demonstrate that the performance of the SaTScan can be influenced by global dependence.

2.1. Conditional Autocorrelated Regressive Model with Gaussian Distribution

For the global dependence model, we will only use the conditional autocorrelated regressive (CAR) model in this study. The CAR model is considered on the basis of its popularity and can be used in spatial regression models. For a CAR model, $\{Z(s_i) : s_i \in D, \ \forall i \in \{1, 2, \ldots, n\}\}$ is defined as a spatial process, or a random process in the spatial domain in lattice D, and $s_i = (u_i, v_i)$ is the location of cell i, where (u_i, v_i) are the coordinates. In practice, s_i is often defined as the geographic center of cell i. Suppose the full conditional distribution of $Z(s_i)$ which follows a Gaussian distribution can be expressed as

$$f(z(s_i)|\{z(s_j) : j \neq i\}) = \frac{1}{\sqrt{2\pi{\sigma_i}^2}} \exp[-\frac{\{z(s_i) - \theta_i(\{z(s_j) : j \neq i\})\}^2}{2{\sigma_i}^2}], \quad (1)$$

where f denotes the conditional density function of $z(s_i)$, and θ_i and σ_i^2 are the conditional mean and variance respectively. The term "pairwise-only dependence" is defined as the condition of θ_i satisfying

$$\theta_i(\{z(s_j) : j \neq i\}) = \mu_i + \sum_{j \neq i} c_{ij}(z(s_j) - \mu_j). \quad (2)$$

Let us assume that the weight of neighborhood information, c_{ij}, is equal to $\rho \times w_{ij}$, where w_{ij} is a known weight and ρ is an unknown spatial dependent parameter. Along with the Hammersley-Clifford theorem, the joint distribution of $\mathbf{Z} \equiv (Z(s_1), \ldots, Z(s_n))^T$, the CAR model, can be established as $\mathbf{Z} \sim Gau(\mu, (I - \rho \times W)^{-1}M)$, where $\mu = (\mu_1, \ldots, \mu_n)^T$ is the mean vector, W is an $n \times n$ matrix whose (i, j) element is w_{ij}, $M \equiv diag({\sigma_1}^2, \ldots, {\sigma_n}^2)$ is an $n \times n$ diagonal matrix, and $(I - \rho \times W)$ is necessarily symmetric and invertible (Besag, 1974; Cressie, 1993).

To estimate the unknown parameters in the CAR model, the maximum likelihood estimates (MLEs) are the most popular method. However, the MLEs do not have the closed forms and this reason could be an intractable problem when we need to do some further computations. Besag (1974) introduced the pseudo-likelihood to solve it and has been proved that the estimates are consistent. We will apply the pseudo-likelihood to obtain the EM estimates in case of clusters which will be discussed in Section 3.

Besides, the selection of a suitable weight function is a difficult problem. In this paper, we will use the "C" type weight function, which is a globally standardized function (the number of cells divided by sums over all number of neighbors in the study region), because it can maintain a symmetric property required for the CAR model. It should be noted that the "spdep" package in the freeware R provides several choices of weight functions; a detailed discussion on the weight functions is provided by Tiefelsdorf et al. (1999).

2.2. Approximation of Conditional Autocorrelated Regressive Model with Poisson Data

For disease clusters, the data are often assumed to be binomial or Poisson distributed instead of being normally distributed. The auto-binomial and auto-Poisson models are established for dealing with these data (Cressie, 1993, Chapter 6). However, these models require extra effort and time in computation to implement autocorrelation structure (such as the CAR model) into generating data and in data analysis.

Instead, it is easier to facilitate the CAR model by a normal approximation through the Freeman-Tukey transformation, as suggested by Cressie and Read (1989). Given that $Z(s_i)$ follows a binomial or Poisson distribution, this transformation is expressed as

$$F_i = (1000 \times Z(s_i)/n_i)^{1/2} + (1000 \times (Z(s_i)+1)/n_i)^{1/2}, \tag{3}$$

where n_i is the population size of cell i. Note that, after this transformation, the covariance matrix will be changed, and it can not be ascertained that the covariance matrix is symmetric and positive definite. We have to do some adjustments to maintain these properties which we will demonstrate them in the application section. Then, we can directly model F_i by the Gaussian CAR model, and all parameters can be estimated in the same way that we introduced for the Gaussian distribution.

In fact, there are other choices to approximate the CAR model. For example, the estimated relative risk, $\hat{r}_i$, can be computed from a Poisson distribution, and the real relative risk r_i can be modeled as a log-linear model,

$$log(r_i) = \theta_i + u_i + \epsilon_i, \tag{4}$$

where u_i is assumed from a normal CAR model and v_i is a random normal effect. This model is usually applied in the Bayesian spatial model (Lawson, 2008). Still, it is more convenient to use the Freeman-Tukey transformation for the sake of consistent variance and fewer parameter estimations.

2.3. The SaTScan

The SaTScan is a type of likelihood ratio test. This test uses a large set of scanning windows to divide the study region into two separated parts and computes the likelihood of cases being in a chosen region divided by that of cases being outside the chosen region. The likelihood ratio statistic is generally expressed as

$$\lambda = \frac{\sup_{Z \in \mathcal{Z}, p>q} L(Z,p,q)}{\sup_{p=q} L(Z,p,q)} = \frac{L(\hat{Z})}{L_0}, \tag{5}$$

where p and q are the incidence rates of inside and outside the chosen region respectively, the $\hat{Z}$ is the region which maximizes λ, and L_0 is the likelihood function under $p = q$; that is, spatial homogeneous (no local cluster). The p-value of Equation (5) can be obtained through the Monte Carlo method (Kulldorff, 1997). In this study, for the case that the data are generated from a CAR model (or an approximation of a CAR model), we adopted the normal distribution frame of the SaTScan, the details of which can be found in Kulldorff et al. (2009).

2.4. Simulation Settings and Performances of the SaTScan in Case of Global Autocorrelation

In this section, we will demonstrate that the global autocorrelation can influence the cluster detection results of the SaTScan. In specific, we shall evaluate the detecting performance of the SaTScan in case of spatial autocorrelation. First, we introduce the simulation setting. Let the study region be a grid area with 20 × 20 squares, and each cell has the same population size $n = 10,000$.

Basically, the log relative risk normal model, which follows Equation (4), is proposed to generate data. Let θ_i is the mean of location i, u_i is assumed from a normal CAR(ρ) model with variance σ_u^2, and ϵ_i is a random normal effect with variance σ_ϵ^2 in Equation (4). To evaluate the effect of global dependence on local clusters, the autocorrelation parameter, ρ, is set as four different values (0, 0.2, 0.5, and 0.8), σ_u^2 is 0.1, and σ_ϵ^2 is 0.01. For the null model, we set all the θ_is to be 0. In addition, suppose the data are obtained from a CAR model with autocorrelation parameter, ρ, the "rooks" type neighbors, and a first-order "C" type neighbor weight function. To avoid edge effects, the "torus" method is applied to construct neighbors. In our synthetic study region, there are 400 cells and each cell has 4 neighbors.

Thus, the "C" type weight is 0.25 (400/1600). The simulations are repeated 100 times in each situation.

It should be noted that we only generate data from above model. However, we will not analyze them in the same model because we focus on finding the local clusters. In our proposed algorithm, only Gaussian distributions are suitable to be applied. So, the Freeman-Tukey transformation is used to convert the simulated data into the Gaussian form, and then the cluster detection algorithm is launched. Here, a normal model and circular windows are set to detect clusters by using the "SaTScan" software, which is free to download from `http://www.satscan.org`.

Although the SaTScan can test if the data are spatially homogeneous, i.e., $H_0 : p = q$ (notations defined in Section 2.3), it cannot distinguish which clustering pattern is (local clusters or global clustering) when the null hypothesis is rejected. In this study, our goal is to check whether the detection method can correctly detect the local clusters when the real spatial pattern is involved with both global clustering and local clusters. For this reason, we define the rejection probability as the probability of rejecting the null hypothesis instead of the type I error, given that there exists global dependence but not the local cluster. Table 1 lists the rejection probability of the SaTScan and it increases with the autocorrelation. It seems that the cluster detection of the SaTScan can be influenced by the global dependence. In addition, the SaTScan is likely to detect more than one cluster as the value ρ increases. In other words, using the SaTScan is possible to detect false clusters when there exists spatial global autocorrelation instead of local clusters.

Table 1. Rejection probability of the circular SaTScan

		Frequencies of Clusters		
ρ	Rejection Probability	0	1	$\geq$2
0	0.02	98	2	0
0.2	0.13	87	13	0
0.5	0.27	73	25	2
0.8	0.61	39	55	6

Notes: The significance level is 0.05. Frequency of clusters represents the number of times that clusters detected by the SaTScan in 100 simulation runs.

Figure 1 shows the frequencies of cells being identified as clustered cells, in which darker areas indicate higher frequencies, and it can be used to explain why the SaTScan can detect false clusters. It is obvious that the false detected cells spread wider in the study region when the autocorrelation is higher. Also, taking the case of $\rho = 0.8$ for example, the darker cells gather around the centered area. Obviously, the rejection probabilities of cells apparently are not randomly and are likely caused by the global autocorrelation.

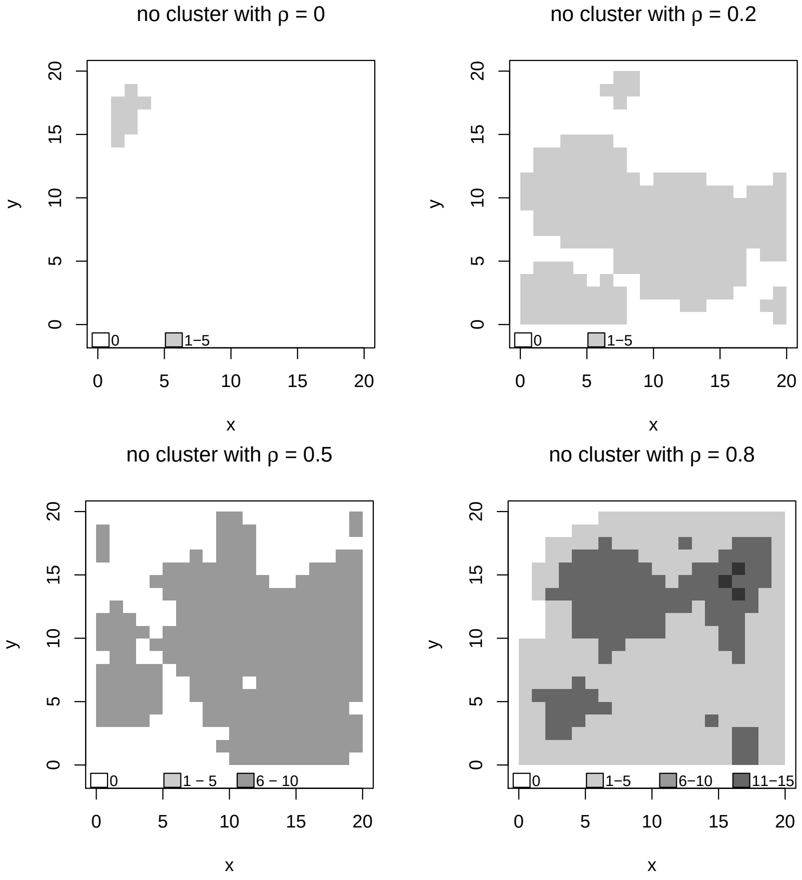

Fig. 1. Images of cluster detection result of the SaTScan in case of spatial autocorrelation without spatial clusters. The labels denote the times of being specified as clustered cells in 100 simulations.

We also check the performance of SaTScan's cluster detection when both autocorrelation and a local cluster are present. Assume that there is a cluster of size 3 × 3 located at the center region. Based on Equation (4), we let the

θ_c as the mean of clustered cells whose values are 0.4, 0.7, and 1, that is, the clustered cells have higher relative risk with mean equal to 1.49, 2.01, and 2.72. In order to provide useful information, some measurements (including the testing power) are used to evaluate the accuracy of detection results. First, we define the terms of true positive, false positive, true negative, and false negative. True positive (TP) cells means the true clustered cells are correctly detected as clusters; false positive (FP) cells means the usual cells are incorrectly detected as clusters; true negative (TN) cells means the usual cells are not identified as clusters; false negative (FN) cells means the true clustered cells are not identified as clusters.

We use the following values to measure the testing performance. First, the power, as its usual definition, is the power to reject the null hypothesis. However, as mentioned in the previous section, the power can not distinguish whether there are local clusters and/or spatial dependence. To check if the power conveys the true information of the local cluster, we define the false alarm as the number of detected clusters which do not include any true positive cell. This measure can show if the identified clusters are real clusters or not. To check if the method can identify all clustered cells, the sensitivity, defined as TP/(TP+FN), is used to measure the proportion of identified clustered cells among all true clustered cells. The positive predictive value (PPV), defined as TP/(TP+FP), is used to measure the proportion of true clustered cells among the identified clustered cells. The specificity is not included in this study since the number of clustered cells is small (9 out of 400) and the specificity is always high no matter what methods are applied.

In Figure 2, we show the preceding measures of the SaTScan under different RR values, and apparently the SaTScan has better performance when the RR becomes larger. However, when there exists stronger global autocorrelation, these measures reveal different information. Take the case of RR = 1.49 ($\theta_c = 0.4$) as an example. As the autocorrelation increases, the power becomes higher but the false alarm goes up as well. This indicates that the SaTScan might detect more than one cluster, similar to those in Table 1. Also, the sensitivity and PPV show that the SaTScan does not have good performances in the cases of the small RR and the autocorrelation worsens the results. In other words, the cluster detection of the SaTScan is obviously influenced by spatial autocorrelation. The accuracy of the SaTScan decreases and the false alarm increases as the autocorrelation increases. This suggests that we shall be cautious about the interpretation of SaTScan's detection result when there exists autocorrelation. However, if the RR is large, spatial autocorrelation does not have a large impact on the performance of the SaTScan with respect to the PPV and the false alarm. For example, for $\rho = 0.8$, the PPV is just 0.1916 for RR = 1.49, raises to 0.6088 for RR = 2.01, and is almost perfect for RR = 2.72. Next, we will investigate the other extreme whether a local cluster can affect the estimation of spatial autocorrelation in a CAR model.

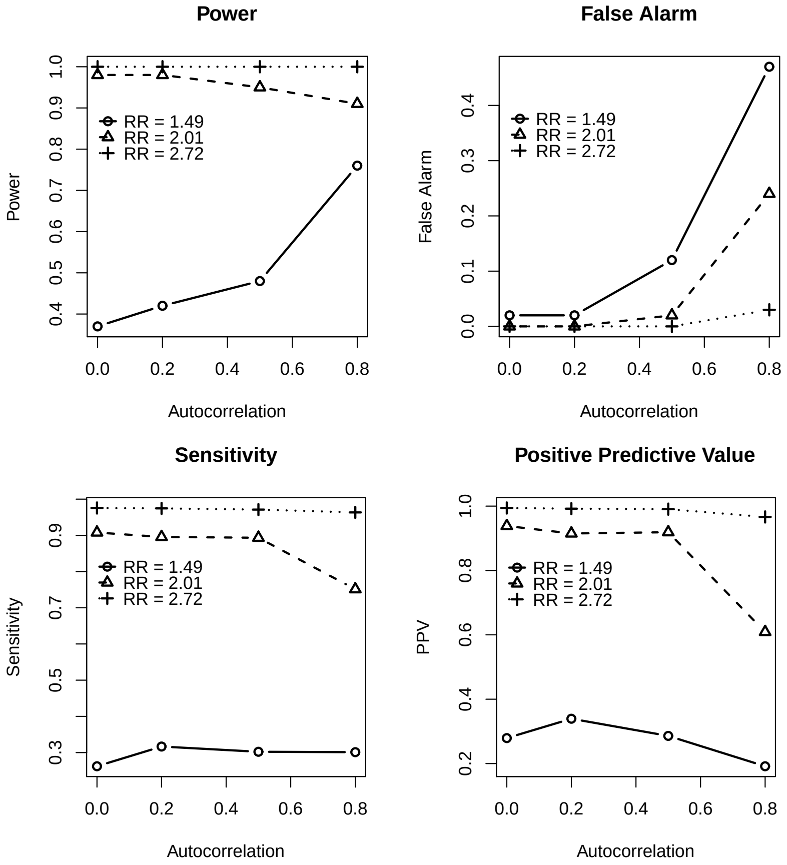

Fig. 2. Results of SaTScan detection in case of autocorrelation and a cluster.

2.5. Performances of a CAR Model when a Cluster is Present

In the previous discussion, we found that the results of SaTScan's cluster detection can be confusing if there exists autocorrelation. We now show that the autocorrelation estimate can also be influenced by the cluster. Suppose a

3 × 3 cluster is located at the center and there is no spatial autocorrelation. Table 2 lists the probability of discovering significant autocorrelation and the average of autocorrelation estimates, with respect to different values of RRs. The RRs are chosen as those in the previous section. The estimates are determined by the "spautolm" function of the package "spdep" and the results are again based on 100 simulation runs.

Table 2. Type-I error of autocorrelation and its estimate

	Relative Risk			
	1.0	1.49	2.01	2.72
Type-I error	0.05	0.12	0.49	0.96
$\hat{\rho}$	-0.0005	0.0462	0.2684	0.5482
S.D. of $\hat{\rho}$	0.1452	0.1681	0.1563	0.1316

Notes: The significance level is 0.05. The values in the table show the average of estimates from 100 replications.

For RR = 1, i.e., no cluster case, the result is acceptable and the average of autocorrelation estimate is close to the true value $\rho = 0$. However, as the RR increases, the type-I error increases significantly and the cluster effect is likely to be mistaken for autocorrelation. We also conducted simulation for the case where both the cluster and autocorrelation exist. The discussion is omitted because the result is similar.

From the simulation results, we found that the spatial cluster and autocorrelation can be confounded to each other. Applying either the SaTScan or CAR model alone can not produce a satisfactory result, and we shall pay attention to the estimations of clusters and autocorrelation separately. Thus, in the next section, we will construct a new approach which is more careful to deal with the local clusters in the spatial pattern involved with global dependence.

3. Proposed Method

In the previous section, it is showed that autocorrelation can affect the performance of SaTScan's cluster detection. We will now introduce an approach that can incorporate global dependence into the cluster detection.

The proposed method can be treated as a combination of spatial models and the SaTScan.

Spatial data possibly consist of both these two effects which are usually confounded. In order to reduce the possibility of mixing two effects, we propose a two-stage approach by handling the autocorrelation first. However, rather than emphasizing the label of one effect, our goal is to make a more flexible space to deal with these two effects. The proposed method can be treated as a kind of step-wise methodologies. It is to estimate the covariance matrix first, and then to scan all the circular regions to find possible clusters, like in the SaTScan. The first step involves estimating spatial autocorrelation by the EM algorithm under the settings of the CAR model. The next step involves generating a series of scanning windows to be the elective clusters, and testing the clusters using the Monte Carlo procedure.

3.1. Pseudo-likelihood

To avoid the intractable term of likelihood in CAR model, we use the pseudo-likelihood (Besag, 1975) to estimate the parameters in the CAR model. Suppose a CAR model is applied to fit a spatial regression model with $\hat{\gamma}_i$s, with $\mu = D\theta$, where D is a designed matrix and θ is the correspondent vector of coefficients, and covariance variance $(I - \rho \times W)^{-1}M$, where W is the neighborhood information weight matrix and $M = V\sigma^2$ is the variance matrix, in which V is a diagonal matrix with diagonal elements being $\{1/N_1, \ldots, 1/N_n\}$ (N_i is the population size in each cell i). In this case, the joint pseudo-likelihood can be expressed as

$$\begin{aligned} p(\Psi) &= \prod_{i=1}^{n} Pr(z(s_i)|z(s_j), j \neq i; \Psi) \\ &= (2\pi\sigma^2/N_i)^{-n/2} \exp[-\sum \frac{N_i}{2\sigma^2}\{z_i - \mu_i - \sum_{j=1}^{n} \rho \times w_{ij}(z_j - \mu_j)\}]. \end{aligned} \quad (6)$$

To obtain the maximum pseudo-likelihood estimates (MPLEs), it is equivalent to solve

$$\begin{aligned} \Omega &= \sum_{i=1}^{n} N_i\{z_i - \mu_i - \sum_{j=1}^{n} \beta_{ij}(z_j - \mu_j)\}^2/2\sigma^2 \\ &= (z - D\theta)^T B^T V^{-1} B(z - D\theta)/2\sigma^2, \end{aligned} \quad (7)$$

where $D\theta = \mu$,$\beta_{ij} = \rho w_{ij}$, $B = (I - \rho W)$, and $V = diag(1/N_1, \ldots, 1/N_n)$.

Thus, the MPLEs can be solved by the ordinary least squares method. Them, the estimates that follow from the above equation can be obtained as

$$\hat{\theta} = (D^T B^T V^{-1} BD)^{-1} D^T B^T V^{-1} BZ, \tag{8}$$
$$\hat{\rho} = (Z - D\theta)^T W^T V^{-1} (Z - D\theta)/\{(Z - D\theta)^T W^T V^{-1} W (Z - D\theta)\}, \tag{9}$$
$$\hat{\sigma}^2 = n^{-1}(Z - D\theta)^T B^T V^{-1} B(Z - D\theta). \tag{10}$$

3.2. EM algorithm

In Section 2, we have known that the estimates can be biased because of the local clusters. The EM algorithm is helpful in estimating the parameters if the locations of clusters are not certain. Besides, it is easy to imped the EM algorithm into the MPLEs because the MPLEs have the closed forms (Equation (8) to (10)).

From the simulation results of the SaTScan, we observed that the SaTScan has relatively consistent sensitivity on detecting clusters but has lower PPV as ρ gets larger, that is, finding more clusters than real ones. For this reason, we treat the clusters detected by the SaTScan as the missing values, and then apply the EM algorithm to estimate the parameters without these missing values.

Suppose the true distribution is $Z \backsim N(D\theta, (I - \phi W)^{-1} V \sigma^2)$. If we divide the vector Z as $(X, Y)^T$, treating X as the observed values and Y as the missing values (outliers/clusters), the joint distribution can be expressed as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \backsim N(\begin{pmatrix} D_x \\ D_y \end{pmatrix} \theta, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}),$$

where D_x and D_y are the designed matrixes of X and Y respectively. Thus, the conditional expectation and variance of Y given X are

$$\mu_{Y|X=x} = D_y\theta + \Sigma_{21}\Sigma_{11}^{-1}(x - D_x\theta), \tag{11}$$
$$Cov(Y|X = x) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{12}$$

Suppose that we have the initial values of $\Psi^{(0)} = (\hat{\theta}^{(0)}, \hat{\rho}^{(0)}, \hat{\sigma^2}^{(0)})^T$, setting $\hat{\theta}^{(0)} = \overline{Z}$, $\hat{\rho}^{(0)} = 0$, and $\hat{\sigma^2}^{(0)} = Var(z)$ in this study. In E-step, according to the Equations (8) to (10), Equations (11) and (12) are used to replace the missing values. Then, based on maximizing the joint pseudo-likelihood

(Equation (6)) the EM estimates can be expressed as

$$\hat{\theta}^{(r+1)} = (D^T {B^{(r)}}^T V^{-1} B^{(r)} D)^{-1} D^T {B^{(r)}}^T V^{-1} B^{(r)} \begin{pmatrix} X \\ \hat{\mu}^{(r)}_{Y|X=x} \end{pmatrix} \tag{13}$$

$$\hat{\rho}^{(r+1)} = \frac{(X - D_x\hat{\theta}^{(r+1)})^T W_{11}(X - D_x\hat{\theta}^{(r+1)}) + tr(W_{22}\hat{\Sigma}^{(r)}_{Y|X})}{(X - D_x\hat{\theta}^{(r+1)})^T U_{11}(X - D_x\hat{\theta}^{(r+1)}) + tr(U_{22}\hat{\Sigma}^{(r)}_{Y|X})} \tag{14}$$

$$\hat{\sigma^2}^{(r+1)} = \frac{1}{n}\{(X - D_x\hat{\theta}^{(r+1)})^T U_{11}(X - D_x\hat{\theta}^{(r+1)}) + tr(U_{22}\hat{\Sigma}_{Y|X})\}, \tag{15}$$

where $B^{(r)} = (I - \rho^{(r)}W)$, and W_{nm} and U_{nm} are the partitions of WV^{-1} and $W^TV^{-1}W$, that is,

$$WV^{-1} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad W^TV^{-1}W = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}.$$

After several iteration steps, the EM estimates will convergence, and then the EM estimates are obtained. It should be noted that if there is no cluster detected by the SaTScan, we will estimate the autocorrelation by the MLE in the CAR model.

3.3. Cluster Model

If the spatial covariance is known, we can directly model the cluster effects and use the generalized least square (GLS) to test their significance. Otherwise, the EM algorithm is used to estimate the spatial covariance. For example, suppose the data are obtained from a CAR model, that is, $Z \sim N(\mu, (I - \hat{\rho}W)^{-1}V\sigma^2)$ after the ρ is estimated. We can estimate μ by the GLS.

To detect clusters, we follow the SaTScan's approach to generate circular windows, which are generated from the centers with different radius, and then compute the test statistics for all the circular window. The main difference between the proposed method and the SaTScan is we take the global dependence into account. Without loss of generality, let the selected window be G and δ_G be the set of the cells which are included in the selected window G. Let $\mu = \mathrm{D}\theta$, where D is an orthogonal matrix with just two indicator columns, one of which contains the cells that belong to the selected region, and the other contains the cells without the selected cells. It can be expressed as

$$\mathbf{D} = (\mathbf{d_0}, \mathbf{d_{\delta_G}}) = \begin{pmatrix} I_{\{s_1 \notin \delta_G\}} & I_{\{s_1 \in \delta_G\}} \\ I_{\{s_2 \notin \delta_G\}} & I_{\{s_2 \in \delta_G\}} \\ \vdots & \vdots \\ I_{\{s_n \notin \delta_G\}} & I_{\{s_n \in \delta_G\}} \end{pmatrix}, \tag{16}$$

where δ_G is a set of the selected cells. $I_{s_k \in \delta_G}$ is an indicator function (1 if $s_k \in \delta_G$ and 0 otherwise), and $I_{s_k \notin \delta_G}$ is an indicator function with the opposite definition. Hence, the problem is transformed to that of choosing a suitable clustered set δ_G. Suppose the coefficients of $\{\mathbf{d_0}, \mathbf{d}_\delta\}$ are $\{\theta_0, \theta_\delta\}$. The hypothesis of testing can be set as $H_0 : \theta_0 = \theta_\delta$ vs. $H_1 : \theta_0 \neq \theta_\delta$, which is exactly defined to test the difference of means of two disjointed sets. Similar to the concept of the geographic analysis machine (Openshaw et al., 1988) and the scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997), we use circular windows to find the optimal clustered set.

To check the cluster effects, we estimate them via the GLS. If the covariance is known or is estimated, we can transform the original spatial dependent data into an independent form. Suppose the spatial covariance matrix is known and non-singular; that is, there is a non-singular matrix P_v satisfying $(I - \hat{\rho}W)^{-1}V = P_v {P_v}^T$. Suppose the original linear model is $Z = D\theta + \xi$, where $\xi \sim N(0, (I - \hat{\rho}W)^{-1}V\sigma^2)$ and the equation can be expressed as $P_v^{-1}Y = P_v^{-1}X\beta + P_v^{-1}\xi$ after the transformation. Let $Z_v = P_v^{-1}Z$, $D_v = P_v^{-1}D$, $\epsilon = P_v^{-1}\xi$, where $\epsilon \sim N(0, \sigma^2)$. Thus, the usual regression model is valid in this case, and the least square estimations of β and σ^2 are $\hat{\theta} = (X_v^T X_v)^{-1} X_v^T Y_v$, and $\hat{\sigma^2} = (Y_v - X_v\theta)^T (Y_v - X_v\theta)/(n-2)$. The testing of $\theta_\delta - \theta_0 \leq 0$ can be conducted via methods used in the regression analysis. Related topics can be found in Rencher (2000). Let $A = (a_0 = -1, a_1 = 1)^T$, $SSH = (A^T\hat{\theta})^T (A^T (X_v^T X_v)^{-1} A)^{-1} (A^T\hat{\theta})$, and $SSE = {Y_v}^T (I - X_v (X_v^T X_v)^{-1} X_v^T)^{-1} Y_v$. Thus, the test statistic is

$$F = \frac{SSH/1}{SSE/(n-2)}. \tag{17}$$

All F scores of all elective windows are recorded, and will be tested via the Monte Carlo method.

It should be noted that the F statistic will be smaller when the autocorrelation gets large. Because the test is equivalent to test the θ in a GLS model with $\Sigma = (I - \rho W)^{-1} V\sigma^2$, the covariance of the θ is $(X^T V^{-1} (I - \rho W) X)^{-1} \sigma^2$. Given that the X and σ^2 are fixed and V is an identical matrix, it is easy to see that the variance will get higher with a larger positive ρ. It is harder to reject the null hypothesis as ρ gets larger.

3.4. Monte Carlo Testing Procedure

To avoid the multiple testing problem in the scanning methods, the Monte Carlo method is applied to our scanning approach. In stead of permuting the disease cases in the study area like what the SaTScan does, we permute the independent residuals from the fitted EM CAR model, and generate the new data sets by the EM estimates. The simulated data are applied in the same algorithm as those in the previous section, and then the maximum F value is recorded. We denote it as F_s for each simulation result where s from 1 to G (simulation times). The procedures are as follows step by step: (1) Obtain the EM estimates for the original data, Z. (2) Obtain the independent residuals as $r = P_v^{-1}(Z - \hat{\theta}_{EM})$ (the notations are defined in Subsection 3.3). (3) Randomly permute r to obtain new residuals r^*. (4) Let $Z^{(s)} = P_v r + \hat{\theta}_{EM}$ be the new data set, and then execute the EM-Scan method procedure to obtain the value F_s.

Suppose we execute the simulations for G times (99 or 999). The actual F scores are compared with these simulated maximum values. Thus, the simulated p-values are obtained as

$$\text{P-value} = \frac{\#\{F_s >= F_{obs}\}_{s=1}^{G} + 1}{G + 1}.$$

The elective clusters are significant if the p-value ≤ 0.05. It should be noted that the significant clusters are possibly overlapped. We will only report the clusters with the largest F value while they are overlapped.

4. Simulations for Method Comparisons

Before we discuss the detection results, it is necessary to check the EM estimates for the Freeman-Tukey transformation data. It should be reminded that the simulation data sets are generated by Equation (4), but the data need to be transformed by the Freeman-Tukey approach before being analyzing. So, it is possible that the model has been biased after the transformation. We will take the transformed data with the clustered effects as the baseline to see if the EM estimates do work. In the following table, we show the MLEs and EM estimates of the spatial autocorrelation ρ and σ^2. The MLEs are obtained by assuming that the locations of the local cluster have known and can be explained by a clustered effect. On the other hand, the EM estimates are computed after treating the clusters detected by the SaTScan as the missing values.

Table 3. Comparisons of MLE and EM estimates

RR	1		1.49		2.01		2.72	
ρ	MLE	EM	MLE	EM	MLE	EM	MLE	EM
0	-0.01	0.00	-0.05	-0.01	-0.04	-0.02	-0.04	-0.02
0.2	0.16	0.16	0.15	0.19	0.15	0.17	0.16	0.18
0.5	0.43	0.43	0.41	0.43	0.40	0.42	0.42	0.43
0.8	0.71	0.66	0.71	0.68	0.72	0.70	0.71	0.71

Notes: For simplification reason, we didn't show the variances of the estimates in the Table. The variances are around 0.11 for all the estimates.

From Table 3, we have observed that the EM estimates are close to the MLEs with the real cluster effect except the cases when RR is lower and $\rho = 0.8$. Because the real simulated model (Equation 4) is a Poisson-Gaussian model, it is not applicable in our approach and we found that the estimates are underestimated from the original settings. Although the Freeman-Tukey transformation can change the original data structure, we think it is easier to find local clusters under the spatial autocorrelation model.

The simulation setting is the same as that in Section 2.4 and we consider both the no-cluster and one-cluster cases. The results are separately listed for the circular SaTScan and the EM-Scan method (EMS). We first compare the probability of rejecting the null hypothesis when there is no cluster. Note that the null hypothesis is all cells having equal disease incidence, and an upper-tail test ($\alpha = 0.05$) is used to check if there are any hot spots. Table 4 shows the simulation results of the two cluster detection methods. The EMS method provides relatively consistent results, but the rejection probability of the SaTScan significantly increases as the autocorrelation gets large. If we use the SaTScan's detection results to judge possible clusters, it might provide distorted information.

Table 4. Rejection probabilities of the cluster detection methods

	Probability of rejecting the null hypothesis with ρ			
Method	0	0.2	0.5	0.8
SaTScan	0.02	0.13	0.27	0.61
EMS	0.06	0.10	0.09	0.17

Notes: The significance level is 0.05.

Because the Monte Carlo procedure of the SaTScan is based on the independent assumption, this assumption can underestimate the variance of the data when a positive autocorrelation exists. The smaller variance could be the reason for raising the false alarms. On the other hand, the proposed method considers the effect of the autocorrelation, and thus reduces the false alarm rates. A similar result appears in the case of a cluster.

Further comparisons can be made in the case of a cluster. The power and the false alarm are shown in Figure 3, where the RR of cluster is equal to 1.49, and 2.01. When RR = 1.49, the two methods obviously have different patterns on the power and the false alarm. The SaTScan has good power, but also has large false alarms. For example, for RR = 1.49 and $\rho = 0.8$, the SaTScan has power 0.76, but with a very high false alarm 0.47. The proposed method is also affected by the larger autocorrelation because the testing variance will get larger when autocorrelation is larger. Therefore, it is more difficult to reject the null hypothesis when the autocorrelation is large. The case where RR = 2.01 shows similar results, and the SaTScan still has the better power but with detecting more false clusters. In contrast, the proposed method does not have a big jump in power or false clusters, when RR increases to 2.01.

Because the power and false alarm can only provide partial information about detection accuracy, we also show the results of sensitivity and PPV, as shown in Figure 4. At the first glance, the sensitivity of the proposed method is not better than that of the SaTScan due to a lower power when RR = 1.49. However, when RR = 2.01, the proposed method is not worse than the SaTScan even the proposed method has a lower power.

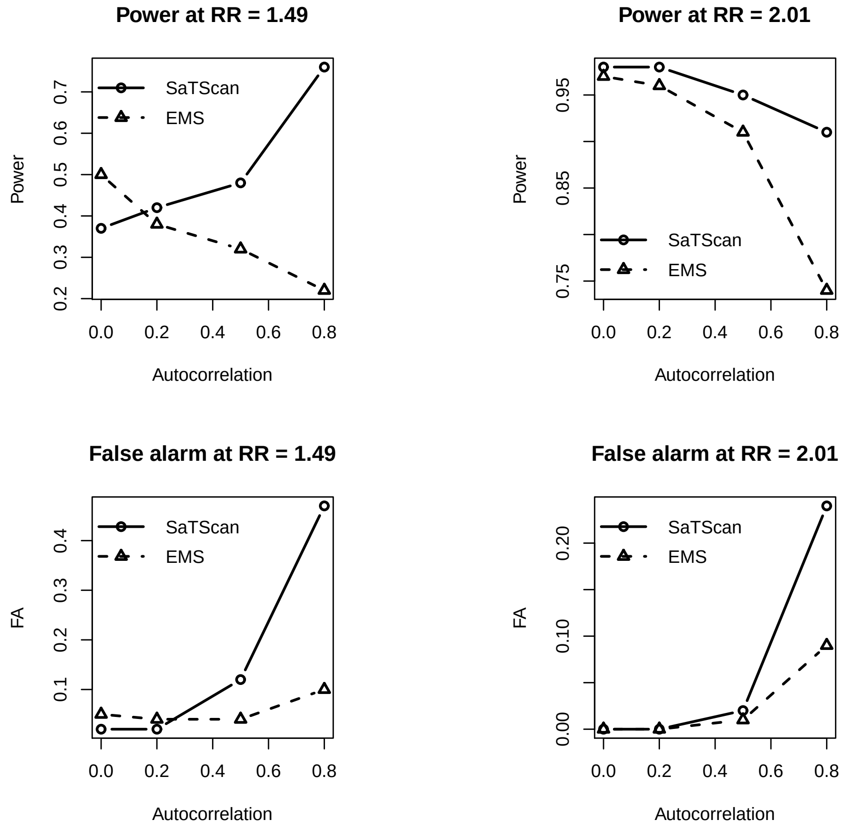

Fig. 3. The power and false alarm comparisons between the SaTScan and the proposed methods with different relative risks.

5. Application

In this section, we use the SIDS data in North Carolina from 1974–1984 to evaluate the proposed EM-Scan method. The data can be found under the package "spdep". In order to compare the results of the proposed method with those of the SaTScan, we use the same setting as Kulldorff (1997). Since the SIDS is a rare disease, we sum up the data from 1974 to 1984 to obtain a sufficiently large sample size to approximate the normal assumption (after the Freeman-Tukey transformation) and Kullforff (1997) also used the same combination. There are 100 counties in North Carolina, and the total number of diseases and live births are 1,503 and 752,354 respectively. The overall average incidence rate is close to 2 per 1,000.

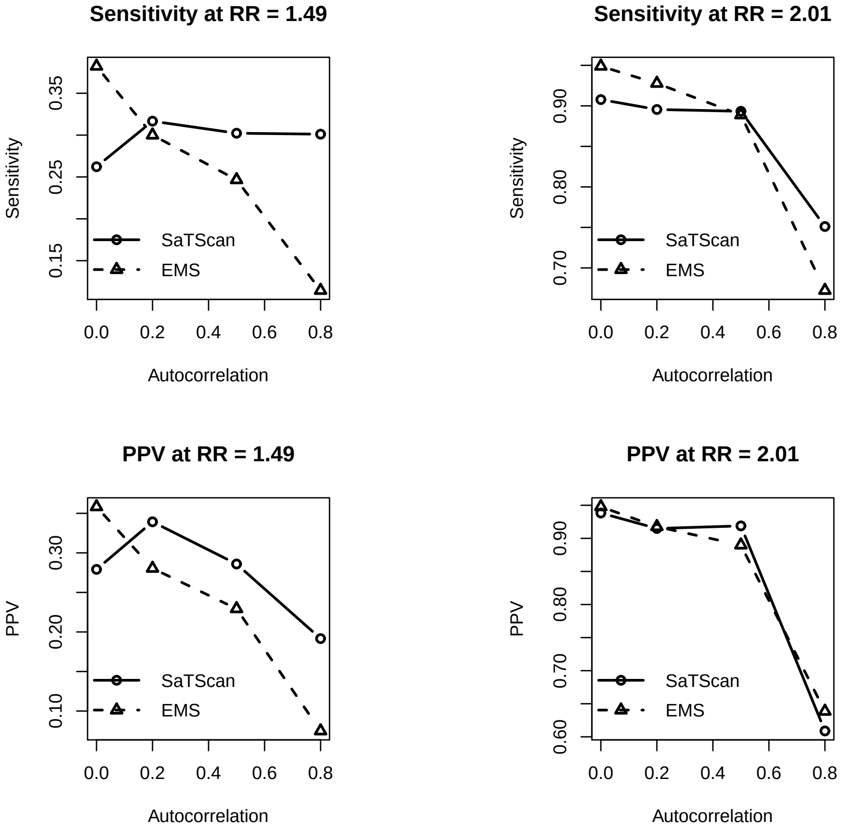

Fig. 4. The sensitivity and PPV comparisons between the SaTScan and the proposed methods with different relative risks.

According to the cases and the population sizes, we can transform the data by the Freeman-Tukey transformation (Figure 5) in order to apply a CAR Gaussian model. Because the neighborhood information is more complicated in the real case, we use the same neighborhood information and weight function suggested by Cressie and Chan (1989), i.e.,

$$w_{ij} = \begin{cases} (m(k)/d_{ij}^k)(n_j/n_i)^{1/2} & j \in N_i, \\ 0 & \text{otherwise}, \end{cases}$$

where d_{ij} is the distance between cells i and j, $m(k)$ is the minimum distance between all cells, n_i and n_j are the population sizes in cells i and j, and the neighbors N_i are defined by a 30 miles criterion, that is, "all those counties with county seat within 30 miles of the seat of the county in question." In this study, we choose $k = 1$ to construct the neighborhood information matrix. In such settings, the covariance of the fitted CAR model is claimed to be symmetric.

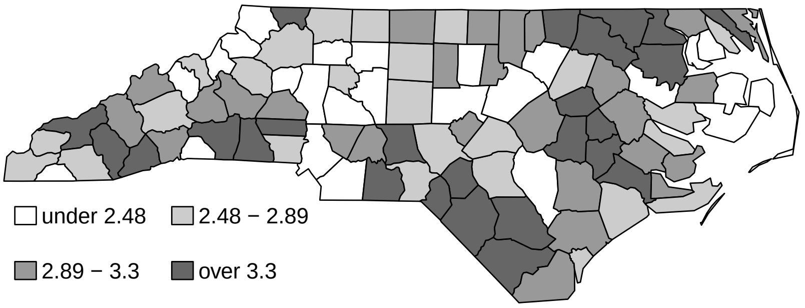

Fig. 5. Map of the 1974–1984 North Carolina SIDS data after the Freeman-Tukey transformation.

The clusters detected by the SaTScan are shown in Figure 6, where we apply the Poisson model to identify SIDS spatial clusters using the software "SaTScan" v9.1.1. Under the significance level 0.05, there are three positive clusters detected by the SaTScan. These results are slightly different than those of Kulldorff (1997) in which only two significantly higher clusters are reported but their locations are not totally the same. The cluster with the single cell in the middle of the study area is the main difference between the detection result of software "SaTScan" and that reported in Kulldorf (1997).

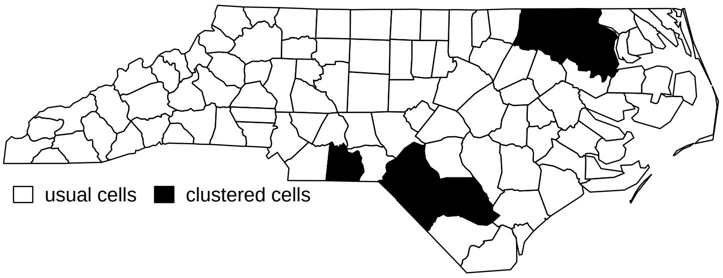

Fig. 6. Clusters detected by the SaTScan.

To apply the proposed EM estimates, the clusters detected by the SaTScan are treated as missing values in advance. We can follow the iterative EM algorithm to obtain the EM estimates. By the EM estimates, the spatial autocorrelation $\hat{\rho} = 0.4884$ is obtained; then, the covariance structure can be treated as a known covariance matrix, and we use the EM-Scan method to detect the clusters. It should be noted that the estimate of the global dependence ρ is 0.8381 if we directly fit the CAR model for the original data without considering the local cluster effects. Based on the simulation Monte Carlo testing in which the EM estimates are treated as the generating parameters, there is no cluster detected under the significance level $\alpha = 0.05$. However, the most significant cluster with p-value = 0.101 (with 999 simulations) as shown in Figure 7.

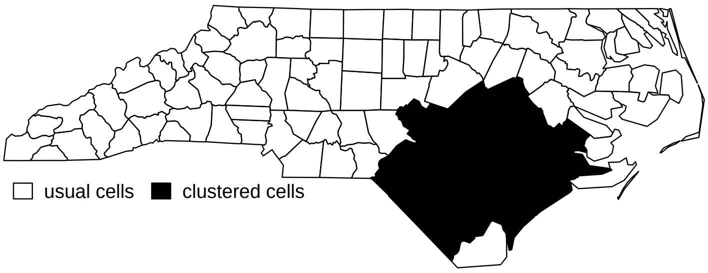

Fig. 7. Cluster detected by the proposed method where the Monte Carlo p-value is 0.101 with 999 simulations.

Although the cluster detected by the proposed method is wider than the clusters detected by the SaTScan, we can not ascertain which clusters are the real ones. Besides, it should be noted that we used different models and data to detect clusters. The clusters detected by the SaTScan are based on the independent Poisson model and using the original data set, but the clusters detected by the proposed method are based on the Freeman-Tukey transformed data and the Gaussian CAR model. The SaTScan would detect a similar cluster as the proposed method if using the Freeman-Tukey transformed data. The analysis of SIDS data also supports that the SaTScan tends to identify more clusters.

In addition, we conduct a residual diagnosis of the proposed method. Although there is no cluster detected by the proposed method, we still use the EM estimates to get the fitted values and the residuals. First, we have to check the normality of the residuals. Based on the Kolmogorov-Smirnov test, the residuals do not reject the normal assumption (p-value = 0.7984).

This means the Freeman-Tukey transformation is proper in analyzing the SIDS data. Also, the spatial dependence is not significant by the Moran's I test (p-value = 0.0839). However, there are 6 outliers detected by the EM fitted values (i.e., z-score ≥ 1.96 or ≤ 1.96). This means there could still be factors affecting the model fitting.

6. Discussions

In this paper, we study the problem of cluster detection when there exists global spatial dependence, that is, a spatial pattern involved both effects. Our goal is to demonstrate that it is difficult to distinguish the global dependence and local clusters. Interestingly, not many past studies focus on discussing this mixed pattern. In this study, we want to show that these two effects can be confounded and need to be taken with care. In particular, we found that the performance of the SaTScan can be influenced by global dependence, and the influence is especially noticeable when the autocorrelation is high.

We propose a two-stage cluster detection method to deal with global dependence. We use computer simulation and empirical data analysis to evaluate the proposed method. The simulation results show that, if the autocorrelation exists, the proposed method can reduce the false alarms but also loses the detection power. Nevertheless, when there is no cluster, the proposed method gives a satisfactory result of detecting fewer false clusters. On the other hand, the SaTScan seems to detect too many false clusters but it also has better powers in detecting clusters. There is still room for improving the cluster detection method when there exists global spatial dependence.

Note that the proposed method and the SaTScan both use Monte Carlo procedure, but the independent assumption for permuting the observations is invalid when there is spatial dependence. Instead, the proposed method uses the Monte Carlo procedure and puts the autocorrelation into account. Because the autocorrelation can raise the variance of the interested estimate (i.e. variance of $\hat{\theta_\delta} - \hat{\theta_0}$), this can help to reduce the false alarm when there is a positive autocorrelation. Although this adjustment sacrifices the power when the RR is small (RR = 1.49), it has similar detection accuracy as the SaTScan when the RR is moderately large (RR = 2.01). Because it seems impossible to completely separate these two effects, the power and accuracy measures may not convey the effectiveness of the proposed model. Two anonymous referees suggest the proposed method can be applied to produce a better map based on the ability of the hybrid (global dependence and local clusters) model. Then, the MSE could be a more reliable measurement to convey the effectiveness of the proposed model. This is a great consideration and we will conduct it in our future researches.

This study mainly argues that global dependence and local clusters are confounded, and it is likely that these two effects can be mislabeled to each other. This is also the reason why we choose the EM algorithm and the two-stage method to estimate the autocorrelation and to identity clusters. However, this approach is possible to lose useful information and to obtain biased estimates, if the missing values are informative or the area identified as missing values is big. One possible modification is to use jackknife estimate, but it has little improvement. We should continue searching other robust methods to acquire stable estimate to the autocorrelation.

On the other hand, we can use other detection methods to choose elective clusters instead of circular windows in the EM-Scan method after handling global dependence. These methods include, for example, the upper level set scan statistic (Patil and Taillie, 2004), the flexible scan statistic (Tango and Takahashi, 2005), and the SaTScan with the elliptic windows (Kulldorff et al., 2006). Then, the rest of the procedures are the same, as described in Section 3. Based on this way, there could be more possible extensions of the proposed method.

7. Acknowledgements

We gratefully acknowledge partial support from Taiwan National Science Council 100-2917-I-004-006. We thank the associate editor and two referees, whose helpful and useful comments improved this article.

References

Anselin L. Local Indicators of Spatial Association-LISA. Geogr Anal 1995;27:93–115.

Besag J. Spatial Interaction and the Statistical Analysis of Lattice Systems. J R Stat Soc Ser B 1974;36:192–236.

Besag J. Statistical analysis of non-lattice data. The Statistician 1975;24:179–195.

Besag J, Newell J. The detection of clusters in rare diseases. J R Stat Soc Ser A 1991;154:143–55.

Cressie N. Statistics for Spatial Data. Rev. ed. New York: Wiley; 1993.

Cressie, N. and Chan, N.H. Spatial modeling of regional variables, J. Amer Statist Assoc 1989;84:393–401.

Cressie N, Read T. Spatial Data Analysis of Regional Counts. Biom J 1989;31:699–719.

Getis A, Ord J. The Analysis of Spatial Association by Use of Distance Statistics. Geogr Anal 1992;24:189–206.

Huang L, Pickle L, Das B. Evaluating Spatial Methods for Investigating Global Clustering and Cluster Detection of Cancer Cases. Stat Med 2008;27:5111–42.

Kulldorff M. A Spatial Scan Statistic. Commun Stat Theory Methods 1997;26:1481–96.

Kulldorff M. Tests of Spatial Randomness Adjusted for an Inhomogeneity J. Amer Statist Assoc 2006;101:1289–1305.

Kulldorff M, Nagarwalla N. Spatial Disease Clusters: Detection and Inference. Stat Med 1995;14:799–810.

Kulldorff M, Tango T, Park P. Power Comparisons for Disease Clustering Tests. Comput Stat Data Anal 2003;42:665–84.

Kulldorff M, Huang L, Pickle L, Duczmal L. An Elliptic Spatial Scan Statistic. Stat Med 2006;25:3929–43.

Kulldorff M, Huang L, Konty K. A Scan Statistic for Continuous Data Based on the Normal Probability model. Int J Health Geogr 2009;8:58.

Lawson A. Disease cluster detection: A critique and a Bayesian proposal. Stat Med 2006;25:897–916.

Lawson A. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Chapman & Hall/CRC; 2008.

Openshaw S, Charlton M, Craft A, Birch J. Investigation of Leukaemia Clusters by Use of a Geographical Analysis Machine. Lancet 1988;331:272–3.

Ord JK, Getis A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr Anal 1995;27:286–306.

Ord JK, Getis A. Testing for local spatial autocorrelation in the presence of global autocorrelation. J Reg Sci 2001;41:411–32.

Patil G, Taillie C. Upper Level Set Scan Statistic for Detecting Arbitrarily Shaped Hotspots. Environ Ecol Stat 2004;11:183–97.

Rencher A. Linear Models in Statistics. New York: Wiley; 2000.

Tango T, Takahashi K. A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters. Int J Health Geogr 2005;4:11.

Takahashi K, Tango T. An Extended Power of Cluster Detection Tests. Stat Med 2006;25:841–52.

Tiefelsdorf M, Griffth D, Boots B. A Variance-stabilizing Coding Scheme for Spatial Link Matrices. Environ Plan A 1999;31:165–80.