Sérgio Henrique Rodrigues Ribeiro

Seleção Ótima dos Parâmetros de Varredura Espacial para Detecção de Conglomerados: Um Estudo de Simulação

Belo Horizonte, 2 de março de 2012

Sérgio Henrique Rodrigues Ribeiro

Seleção Ótima dos Parâmetros de Varredura Espacial para Detecção de Conglomerados: Um Estudo de Simulação

Dissertação apresentada como requisito parcial para obtenção de grau de Mestre em Estatística pela Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Marcelo Azevedo Costa

Programa de Pós-Graduação em Estatística
Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, 2 de março de 2012

Sérgio Henrique Rodrigues Ribeiro

Seleção Ótima dos Parâmetros de Varredura Espacial para Detecção de Conglomerados: Um Estudo de Simulação

Esta dissertação foi julgada adequada à obtenção do título de Mestre em Estatística e aprovada em sua forma final pelo Curso de Mestrado em Estatística da Universidade Federal de Minas Gerais.

Belo Horizonte, 2 de março de 2012

_____

Prof. Dr. Marcelo Azevedo Costa
Universidade Federal de Minas Gerais

_____

Prof. Dr. Renato Martins Assunção
Universidade Federal de Minas Gerais

_____

Prof. Dr. Alexandre Loureiros Rodrigues
Universidade Federal do Espírito Santo

Agradecimentos

Agradeço a Deus por mais essa conquista. Por todas as bênçãos que vejo e por todas as bênçãos que os nossos olhos não veem.

À minha amada esposa Andréa por ter acreditado, pela dedicação e paciência e, acima de tudo, por te me amado desde o princípio.

Agradeço a minha amada mãe Laura, meu pai Sérgio (in memoriam), meu padrasto João, as minhas irmãs Patrícia e Cristina pelo apoio, de todas as formas possíveis, para que eu chegasse até aqui.

Aos meus sobrinhos Sabine, Ana e Arthur pelos momentos de carinho e descontração que me trazem tanta felicidade e renovo.

Ao meu professor orientador Dr. Marcelo A. Costa pela parceria e confiança desde a minha graduação. Pelos imensuráveis ensinamentos, direcionamentos, ideias e receptíveis diálogos.

Aos professores do curso de graduação e pós-graduação em estatística por todo o conhecimento e experiência compartilhada. Aos funcionários e todo o corpo docente do Departamento de Estatística pelo esforço e trabalho.

Aos meus amigos estatísticos que, seja na graduação ou pós-graduação, enriqueceram o dia-a-dia com conhecimento e alegria. Que o auxílio mútuo na estatística e na vida dure por longos anos.

Aos queridos velhos amigos pelo incentivo e paciência.

"*Agora, pois, permanecem a fé, a esperança e o amor, estes três; mas o maior destes é o amor."* 1 Coríntios 13:13

"*At any rate, I am convinced that He does not play dice.*"

Albert Einstein se referindo a Deus em uma carta para Max Born, 4 de Dezembro de 1926: 'Einstein und Born Briefwechsel' (1969) p. 130.

Resumo

O desenvolvimento e o estudo de métodos que sejam capazes de detectar eficientemente um conjunto de áreas que tenha uma maior ou menor incidência de um determinado evento são de extrema importância para a sociedade. Por exemplo, um método que indique a existência de um conjunto de bairros que tenha uma maior incidência de crimes ou casos de dengue de uma cidade. Nesse contexto nos deparamos com os métodos de análises estatísticas de conglomerados espaciais. Carpenter (2011) relata que uma das principais razões para o grande uso das análises estatísticas de conglomerados espaciais é a disponibilização gratuita de softwares, como o SaTScan. Por exemplo, no site do SaTScan (www.satscan.org; acessado em 1 de janeiro de 2011) pode ser encontrada uma lista de aplicações nas áreas de doenças infecciosas, parasitologia, vigilância síndrômica, câncer, pediatria, geriatria, doenças neurológicas, psicologia, demografia, veterinária, botânica, silvicultura, ecologia e meio-ambiente, desastres naturais, criminologia, transporte, entre outras. Costa e Kulldorff (2009) revisaram algumas dessas aplicações.

A estatística de varredura espacial é baseada no clássico artigo de Naus (1965) que desenvolve expressões matemáticas para calcular a probabilidade de encontrar k pontos dentro de uma janela retangular fixa que varre a área de um quadrado unitário. Kulldorff (1997), com a estatística de varredura circular, estendeu esta abordagem assumindo um processo de Poisson ou Bernoulli que gera eventos ou casos em uma região geográfica, também conhecida como região de estudo. A região de estudo pode ser dividida em áreas menores onde as populações sob risco e os casos são observados. Os candidatos a conglomerados são gerados por círculos com os seus centros posicionados nos centroides das áreas e, em seguida, variando os seus raios. Para cada círculo, uma estatística da razão da log-verossimilhança é calculada. O círculo com a maior estatística observada é a potencial sub-região crítica. A inferência sob a suposição de aleatoriedade espacial é realizada por meio de simulações Monte Carlo (Dwass, 1957; Turnbull et al., 1990).

Algumas extensões e modificações da estatística de varredura puramente espacial são encontradas na literatura. Por exemplo, a estatística de varredura puramente espacial foi estendida para três dimensões (Kulldorff et al., 1998; Kulldorff, 2001;

Kulldorff et al., 2005), onde o tempo geralmente representa a terceira dimensão. Novos modelos de probabilidade também foram propostos (Jung et al., 2010, 2007; Huang et al., 2007; Kulldorff et al., 2009; Huang et al., 2009), bem como geometrias diferentes para a forma do conglomerado (Duczmal e Assunção, 2004; Patil e Taillie, 2004; Assunção et al., 2006; Takahashi et al., 2005, 2008; Costa et al., 2012). No entanto, a estatística de varredura circular e elíptica (Kulldorff et al., 2006) são as metodologias mais utilizadas. Pode-se argumentar que as estatísticas de varredura circular e elíptica são amplamente utilizadas porque estão disponíveis no software SaTScan.

O usuário da estatística de varredura circular precisa selecionar apenas um parâmetro, que é o tamanho máximo da janela de varredura. Este parâmetro é normalmente escolhido em termos percentuais da população total na região de estudo. Alguns autores, Kulldorff e Nagarwalla (1995), Coulston e Riitters (2003), Forand et al. (2002), Donnan et al. (2005) e Chaput et al. (2002), discutem a utilização de alguns valores específicos desse parâmetro. Costa e Kulldorff (2009) descrevem algumas razões para escolher um tamanho menor de conglomerado. Uma razão é que um conglomerado circular menor pode ser um indício de que o conglomerado verdadeiro tem uma forma irregular, assim, representa uma análise exploratória para conglomerados irregulares, antes da execução de qualquer método de detecção de conglomerados irregulares. Outra razão, é que a escolha por um tamanho de conglomerado menor pode ser feita com base nos recursos disponíveis para a intervenção, que é o caso em estudos epidemiológicos e aplicações de vigilância sindrômica. Por exemplo, a detecção de um conglomerado com um tamanho de 50% do território nacional dos EUA não é muito informativa.

Neste trabalho, são estudas as medidas de desempenho dos métodos de detecção de conglomerados espaciais para diferentes valores do tamanho máximo do conglomerado com base em cenários simulados. Procuramos evidências empíricas sobre valor ideal para o parâmetro do tamanho máximo do conglomerado. Simulações extensivas foram feitas utilizando a estatística de varredura circular, a estatística de varredura elíptica e a estatística de varredura double (Costa et al., 2012). Esta última aplica uma regra de parada prematura ao processo de construção do candidato à conglomerado, o que pode evitar que o conglomerado detectado alcance o tamanho máximo previamente escolhido pelo usuário. Foram investigados conglomerados simulados de geometria circular e irregular. Além disso, foram investigados os

conglomerados secundários detectados, sob a suposição de que a estatística de varredura espacial pode dividir um único conglomerado desconhecido em vários pedaços, ou seja, detectar conglomerados primários e secundários.

Os resultados indicam que existem escolhas ótimas e únicas para cada um dos seguintes parâmetros: o tamanho máximo do conglomerado, a geometria, e o critério de sobreposição entre os conglomerados primários e secundários. Estas escolhas otimizam as medidas de desempenho: sensibilidade, poder, especificidade e erro de classificação. Contudo, não há uma escolha única, entre todos os parâmetros, que forneça os melhores resultados para todas as medidas de desempenho avaliadas. Os detalhes são apresentados a seguir.

Optimal Selection of the Spatial Scan Parameters for Cluster Detection: A Simulation Study

Sérgio Henrique Rodrigues Ribeiro[a], Marcelo Azevedo Costa[a,1,]

[a]*Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil*

Abstract

Circular and elliptic spatial scan statistics requires the user to choose a maximum cluster size. A common value for this parameter is 50% of the underlying population. In addition to the detected primary cluster, the user may be interested in the analysis of significant secondary clusters. It can also be argued that if the true cluster is irregular, then choosing a small value for the maximum cluster size and evaluating significant secondary clusters may improve cluster detection and avoid the use of irregular cluster methods. This work explores the performance of the circular, elliptic and double scan statistics for different values of the maximum cluster size and different options for the analysis of secondary clusters. Empirical results show that for hot-spot clusters, the analysis of secondary clusters which are statistically significant do not improve the detection of the true unknown cluster, on average. There is evidence that a variable maximum cluster size improves performance. That is, the double scan statistic applies an early-stopping procedure which improves positive predictive values.

Keywords: spatial scan statistic, simulation study.

1. Introduction

Carpenter (2011) reports that one of the main reasons why statistical spatial cluster analysis has been widely used is the availability of free spatial softwares, such as SaTScan. For instance, in SaTScan's web site (`www.satscan.`

Email address: `azevedo@est.ufmg.br` (Marcelo Azevedo Costa)

[1]corresponding author

org; accessed January 1, 2011) one can find a selected list of applications in the fields of infectious diseases, parasitology, syndromic surveillance, cancer, cardiology, rheumatology/auto-immune diseases, among others. Costa and Kulldorff (2009) review some of these applications and find some very interesting patterns among the fields of applications. Among the findings, Costa and Kulldorff (2009) report the use of a varying maximum cluster size for the circular scan statistic in order to improve cluster detection.

The purely spatial scan statistic is based on the classical paper of Naus (1965) which develops mathematical expressions to calculate the probability of finding k points inside a fixed rectangular window moving along a unit square. Kulldorff (1997) extended this approach by assuming a Poisson or Bernoulli process which generates events or cases in a geographical region, also known as the study region. The study region can be further divided into smaller disjoint subregions (or areas) where populations at risk and cases are observed. Circular cluster candidates are created by centering circles at the centroids of the areas and then varying their radii. For each circle, a log-likelihood ratio statistic is calculated. The circle with the largest observed statistic is a potential critical subregion. Inference under the assumption of spatial randomness is carried out using Monte Carlo simulations (Dwass, 1957; Turnbull et al., 1990).

The circular scan statistic requires the user to select only one parameter, which is the maximum window size. This parameter is usually chosen in terms of the percentage of the total population in the study region. For example, a maximum size of 30% means that the radii of the circles will increase until reaching 30% of the total population inside the circles. Kulldorff and Nagarwalla (1995) claim that this value should be chosen as 50% in order to avoid 'negative clusters', which are a few areas with low incidence rates outside the circle. Nevertheless, many applications use smaller values for practical reasons. For example, Coulston and Riitters (2003) choose a maximum cluster size of 20% due to forestland discontinuity, and later they run the circular scan using the previously detected cluster as the new study region and report "several small clusters arranged in a linear fashion along Interstate 95". Forand et al. (2002) restrict the number of cases inside the cluster candidate to be no more than 2.5% of all cases to better focus on geographic areas covered by the participant hospitals where the data were collected. Donnan et al. (2005) reduce the maximum cluster size to check for smaller but significant clusters. Chaput et al. (2002) use maximum cluster sizes of 50% and 25%, arguing that "smaller size allows detecting more com-

pact cluster, which indeed generated a primary significant cluster with higher relative risk". Costa and Kulldorff (2009) describe some reasons for choosing a smaller cluster size. One reason is that a smaller circular cluster size can indicate whether the true cluster has an irregular shape. Therefore, choosing a smaller cluster size represents an exploratory analysis for irregular clusters, before running any irregular cluster detection method. Another reason is that a smaller cluster size may be chosen based on available resources for intervention, which is the case in epidemiological studies and disease surveillance applications. However, the effects of different values of maximum cluster size on cluster analysis performance have not been properly investigated.

Regarding circular and irregular shapes, the literature shows that for truly circular clusters, the circular scan statistic achieves better performance whereas for truly irregular clusters, irregular methods achieves better performance, as expected. These results are usually based only on the analysis of the significant primary detected clusters. That is, by means of simulations and given the primary detected cluster, the location and population of the detected cluster are compared to the location and population of the true cluster. Nevertheless, the circular scan statistic also reports secondary clusters. Thus, it can be argued that by selecting properly the maximum cluster size parameter and looking at the secondary clusters, then circular and irregular clusters are found. If so, then detection results are to be close to the elliptic (Kulldorff et al., 2006) or irregularly cluster methods (Duczmal and Assunção, 2004; Patil and Taillie, 2004; Assunção et al., 2006; Takahashi et al., 2005, 2008; Costa et al., 2012).

In this work we investigate possible choices for the maximum cluster size parameter. We study performance measures for spatial clustering for different values of the maximum cluster size using simulated scenarios. We use power, sensitivity, positive predictive value and misclassification statistics in order to provide further insights into the effectiveness of different parameterizations in SaTScan. We search for empirical evidence about the optimum maximum cluster size. Extensive simulation is provided for the circular scan statistic, elliptic scan statistic and the novel double connected scan statistic. The latter applies an early-stopping rule, which may avoid detected clusters with the maximum size, previously chosen by the user. Simulated irregular and circular shapes are investigated. In addition, secondary clusters are also investigated, under the assumption that the spatial scan statistics may split a unique unknown cluster into detected primary and secondary clusters.

Results indicate that there are optimal and unique choices for each of the

following parameters: the maximum cluster size, geometry, and the overlapping criterion among primary and secondary clusters. These choices optimize the following performance measures: power, sensitivity, positive predictive value, and misclassification. For instance, a maximum cluster size of 50% provides better power and sensitivity. The double scan statistic provides better positive predictive values, although a maximum cluster size of 5% may also achieve good results. There is no single choice, among all the parameters, that provides the best results for all evaluated performance measures. In general, the analysis of secondary clusters do not improve cluster analysis, except for the sensitivity statistic, which is improved.

This paper is organized as follows. The next section describes the circular, elliptic and double connected scan statistics. Following is a description of the performance measures (section 3). Section 4 presents the simulation study. Section 5 presents the results and the discussion, and the conclusion is provided in section 6.

2. Spatial Scan statistics

2.1. The circular scan statistic

The spatial scan statistic proposed by Kulldorff (1997) uses a circular window to scan a geographical region. The methodology is presented as follows: let A be the geographical region, also named as the study region, partitioned into K disjoint subregions (or areas), for example, counties or states. Let c_i and n_i be the number of cases and populations at risk of area i. Under the null hypothesis of spatial randomness, the cases are uniformly distributed in the population. Therefore, the number of cases in the ith area is Poisson distributed with the expected number of cases, μ_i, proportional to its population,

$$H_0 : c_i \sim Poisson(\mu_i = \lambda n_i)$$

Under the null hypothesis $\hat{\lambda} = C/N$, where C is the total number of cases and N is the total population. Under the alternative hypothesis, there is one spatial cluster at an unknown location. Define Z as the set of all possible circular clusters z. For each cluster z let c_z and n_z be the number of cases and populations inside cluster z. Using a Poisson model, the likelihood ratio test statistic associated with the most likely cluster is given by:

$$\frac{L(\hat{z},\hat{p},\hat{r})}{L_0} = \sup_z \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z} \quad (1)$$

where μ_z is the expected number of cases under the null hypothesis, $\mu_z = C \cdot n_z/N$.

In sequence, Monte Carlo simulations (Dwass, 1957) are applied to address the statistical significance of the most likely cluster. Under the null hypothesis, C simulated cases are assigned to areas using a multinomial distribution and the likelihood ratio test statistic is computed. This procedure is repeated many times to generate an empirical distribution of the likelihood test statistic under the null hypothesis. Finally, the p-value is produced by comparing the observed likelihood ratio test statistic to its empirical distribution. Further details can be found in Kulldorff (1997) or in Costa et al. (2012).

2.2. The elliptic scan statistic

The first difference between the circular scan statistic and the elliptic scan statistic (Kulldorff et al., 2006) is the replacement of the set Z of circular clusters by a set of elliptical clusters. An ellipse is uniquely defined by the coordinates of its centroid, shape, angle and size. "The shape is the ratio of the longest to the shortest axis of the ellipse and the angle θ is the angle between the horizontal line and the semimajor axis of the ellipse" (Kulldorff et al., 2006). Circles are ellipses with shapes equal to one. For mostly computational reasons, restricted values are used for shape and angle. For instance, SaTScan default values for the number of angles are 4, 6, 9, 12 and 15, and for the shapes are 1.5, 2, 3, 4 and 5. Values for the angles are evenly chosen around the circle, and $\theta = \frac{\pi}{2}$ is always included.

The second difference is a non-compactness penalty parameter of the form $[4s/(s+1)^2]^a$ which multiplies the log likelihood ratio statistic. s is the shape parameter and a is the non-compactness penalty parameter. Standard values for the parameter a are: $a = 1$ (strong penalty), $a = \frac{1}{2}$ (medium penalty) and $a = 0$ (no penalty). Medium to strong penalty values are generally used to penalize long and narrow ellipses, because these clusters might have geographically disconnected areas and large values of the likelihood ratio test statistic.

2.3. The double connected scan statistic

The double connected scan statistic, hereafter named double scan statistic (Costa et al., 2012), belongs to the group of scan statistics with irregular geometry (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Takahashi et al., 2005; Assunção et al., 2006; Duczmal et al., 2007; Takahashi et al., 2008). It uses the geographical adjacency information to create an interconnected graph structure among the areas. From this graph structure, candidate clusters are created. Initially, the algorithm starts the cluster with one area and evaluates the likelihood ratio statistic for the neighboring areas which are directly connected to the first one. The neighboring area that increases the likelihood ratio statistic the most is definitely aggregated into the cluster. Otherwise, if there is no neighboring area that increases the likelihood ratio statistic, then the neighboring area that decreases the least the likelihood ratio statistic is definitely aggregated into the cluster. Therefore, the cluster now has two areas. From this point forward, the algorithm selects only the neighboring areas of the cluster which are connected to al least two areas inside the cluster. Among these areas, the area that most increases the likelihood ratio statistic is aggregated into the cluster. If the neighboring areas do not increase the likelihood ratio statistic, then the algorithm stops and starts a new cluster from a new area. Otherwise, the growing process continues to evaluate the likelihood ratio statistic of neighboring areas that are connected to at least two areas already inside the cluster. The growing process stops when the maximum cluster size is reached, or if there are no neighboring areas either double connected or incapable of increasing the current likelihood ratio statistic of the cluster. See Costa et al. (2012) for further details.

Different from the elliptic scan statistic, which requires the user to select shape, angle and penalty parameters, the double scan statistic requires only the maximum cluster size. Therefore, it requires the same parameter as the standard circular scan statistic. Moreover, by applying the graph structure to create cluster candidates, the method provides more flexibility to the cluster shape. The double connected criterion provides a subtle but effective non-parametric penalty to the shape of the cluster and to the likelihood ratio statistic. Therefore, it avoids finding non-informative clusters which are extremely large, too irregularly shaped and have high relative risk and high value of the likelihood ratio statistic.

The double scan statistic also requires a maximum cluster size parameter. Nevertheless, created clusters may not reach the maximum size. This is

because the double scan statistic applies an early-stopping procedure, that is, the growing process stops if there are no neighboring areas able to increase the likelihood ratio statistic of the current cluster. This procedure creates new dynamics in cluster set Z. Basically, this approach is more data driven and it usually generates smaller clusters. This characteristic improves some statistical performances of the method and it is further explored in the simulation study.

3. Performance Measures

As stated by Read et al. (2011), in order to quantify spatial accuracy of the spatial scan statistic, "the literature presents a patchwork of different measures and nomenclatures". The statistical power is a very common performance measure, see for instance Kulldorff et al. (2003, 2004, 2006); Duczmal et al. (2006); Song and Kulldorff (2003); Costa et al. (2012), among others. In order to measure the differences between true and detected clusters, the sensitivity (Huang et al., 2007; Que et al., 2008; Costa and Assunção, 2005; Costa et al., 2012) and positive predictive values (Takahashi et al., 2005; Jung et al., 2010; Que et al., 2008; Costa et al., 2012) are also commonly used.

Following Costa et al. (2012) we use four evaluation metrics: estimated power, sensitivity, positive predictive value (PPV) and misclassification. The sensitivity, PPV and misclassification statistics are estimated in terms of the proportion of the true cluster population,the detected cluster population, and the total population, respectively. Previous studies use the number of areas as the basis for calculating the performance measures. Costa et al. (2012) claim that using the population as the basis provides more robust estimates.

In this work, the estimated power is the proportion of p-values smaller or equal to 0.05 in $10,000$ simulations for each cluster model. The cluster models are shown in section 4.1.

Let $\hat{z}$ be the population inside the detected cluster and z be the population inside the true cluster. The sensitivity represents the proportion of the population in the true cluster that is part of the detected cluster, or $n(z \cap \hat{z})/n(z)$. If the sensitivity is one, then the detected cluster completely contains the true cluster. Nevertheless, a large detected cluster may contain a smaller true cluster and, in this situation, the sensitivity statistic is also unitary.

The PPV is the proportion of the detected cluster population that is part of the true cluster population, or $n(z \cap \hat{z})/n(\hat{z})$. If the PPV is one, then the detected cluster contains only elements of the true cluster. That is, a smaller detected cluster that contains only areas of a large true cluster has PPV equal to one. The optimal detected cluster has sensitivity and specificity equal to one.

The misclassification is the percentage of the total population that belongs to either the detected or true cluster but does not lie in the intersection between the two, or $n[(z \cup \hat{z}) \cap (z \cap \hat{z})^c]/N$. Therefore, it accounts for both the detected cluster populations which do not belong to the true cluster, and for the true cluster populations which do not lie in the detected cluster. The optimal value for misclassification is zero, therefore spatial scan statistics present good performance if they achieve misclassification results as close to zero as possible.

4. Simulation study

4.1. Cluster Models

Two benchmark data sets were used to compare the circular, elliptic and *double* scan statistics. The first data set comprises circular cluster models (Kulldorff et al., 2003). The underlying population is the female population in the Northeastern USA from the 1990 census. The region under study has 245 counties in Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland, and the District of Columbia. The total population is 29,535,210 individuals. The map of the population is shown in Figure 1. We used nine different circular cluster models with 1, 4 and 16 counties located in mixed, rural and urban areas, as shown in Figure 2 (a). Figure 2 (a) shows the rural, mixed and urban cluster models with size 16. Circular cluster models with size 1 use only the county in the center of the circle, whereas cluster models with size 4 use the counties closer to the center of the circles. This data set has previously been used to evaluate spatial scan statistics such as the circular scan (Kulldorff et al., 2003, 2004; Costa and Assunção, 2005), elliptic scan (Kulldorff et al., 2006) and irregularly-shaped scan statistics (Duczmal et al., 2006; Costa et al., 2012), and it is available at `http://www.satscan.org/`. The circular cluster models are shown in Figure 2 (a).

The second data set comprises irregular cluster models (Duczmal et al., 2006) and it uses the same geographical region, underlying population and

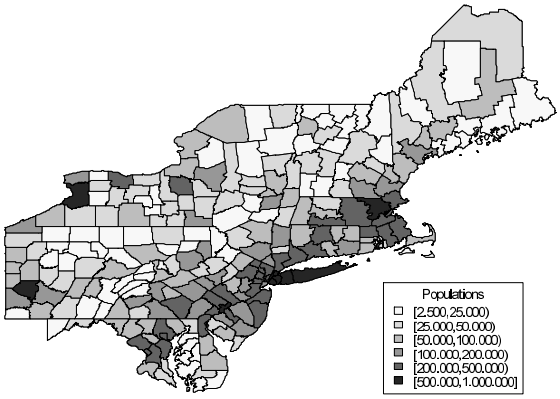


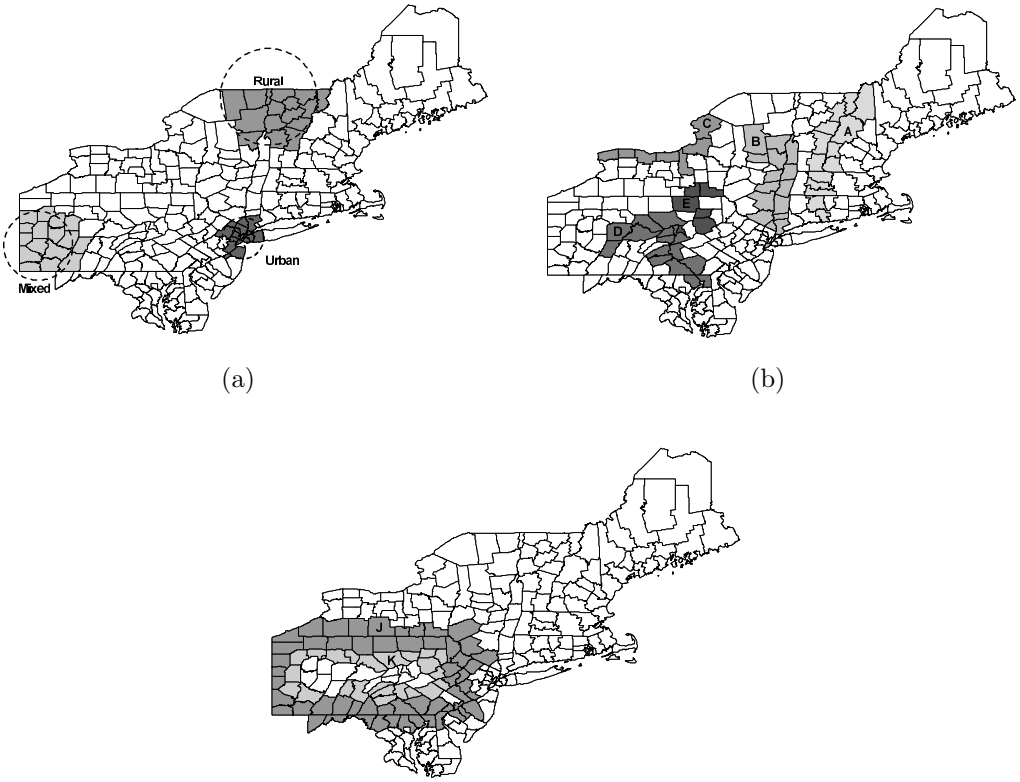
Figure 1: Underlying population in the Northeastern U.S. data set.

number of cases as in the circular cluster models. We use seven different cluster models, containing between 7 and 78 counties, as shown in Figures 2 (b) and 2 (c). Some of the irregular shapes are based on landscape features (Connecticut river, Hudson river, Lake Ontario Coast, among others), while others are based on political boundaries (Pennsylvania Internal/External Border, Pennsylvania Sub-Internal/Internal/External Border).

For both circular and irregular cluster models, 600 cases are randomly distributed across the region in proportion to the population, except within the cluster areas that have a higher relative risk. Relative risks were chosen so that the null hypothesis would be rejected with probability 0.999 when using a standard test for the difference between two binomial proportions and assuming that the cluster location is known. The relative risks in the cluster models are shown in Table 1. Table 1 also shows the number of counties in each cluster model, the populations inside the cluster models and the percentage of the population in the cluster models compared to the total population in the study region. For each one of the circular and irregular cluster models, a total of 10,000 simulated replicas were available. 9,999 Monte Carlo runs for each test statistic were used to estimate the associated *p-value*.

4.2. The spatial scan parameters

We evaluated the circular and elliptical scan statistics with maximum cluster sizes of 2%, 5%, 10%, 15%, 25% and 50% of the total population. SaTScan also stores significant secondary clusters "as if there were no other



(a) (b)

(c)

Figure 2: Circular cluster models in rural, mixed and urban regions (a). Irregular cluster models A, B, C, D and E (b). Irregular cluster models J and K (c). Cluster K includes the elements of cluster J

clusters in the data set" (Kulldorff, 2010). The user can set to what extent overlapping clusters are reported in the results files. Six different options are available: (1) secondary clusters do not overlap with a previously reported cluster; (2) secondary clusters are not centered in a previously reported cluster and do not contain the center of a previously reported cluster; (3) secondary clusters are not centered in a previously reported cluster; (4) secondary clusters do not contain the center of a previously reported cluster; (5) secondary clusters are not centered in a previously reported cluster that contains the center of a previously reported cluster; and (6) no restrictions. See Kulldorff (2010) for further details. We include analyses of secondary clusters using options (1) and (6), along with no consideration of secondary clusters. We do so because these are the common options explored in the literature. We include the areas of the secondary cluster into the primary cluster as if only one cluster were detected. The overlapping areas were included into the primary cluster only one time. We choose an α-level of 0.05 (5%) to report statistically significant secondary clusters.

For the elliptic scan statistics we apply SaTScan's default values for the shape parameter and for the angle parameter, as previously described. We choose the non-compactness penalty parameter as $a = 0$ because a strong penalty ($a = 1$) generates clusters which are very similar to the circular clusters, and a medium penalty ($a = \frac{1}{2}$) does not improve cluster detection performances, as reported in Costa et al. (2012).

4.3. Averaged performance measures

Different from Costa et al. (2012), that calculates the sensitivity, specificity and misclassification for all simulated cluster, despite their significance, we first test whether each detected cluster is significant. If the detected cluster is not significant then the detected population is set as zero, otherwise the detected population is not changed. The performance measures are calculated for both cases, whether there is a significant cluster or not; but, in the latter case, the detected population is zero. We report the mean values over the 10,000 replicas for each cluster model, for each value of the maximum cluster size, and for the circular, elliptic and double scan statistics.

5. Results

To provide a visual comparison of the results, the maximum values of power, sensitivity and PPV for each cluster model, i.e., for each row, are

written in boldface (see Tables 2, 3 and 4). For misclassification, the minimum values are also in boldface (see Table 5). For each row, the cells whose values are no greater than of 0.01 (1%) distance from the boldface value are shaded in dark gray. The cells whose values are within 0.01 and 0.02 (1% - 2%) distance from the boldface value, are shaded in light gray. By doing so, it becomes easier to compare values which are closer to the best result. The results for the double scan statistic are replicated for both circular and elliptic scan statistics.

Table 2 shows the power results. In general, best values of power (in boldface) are achieved for any maximum cluster size except for 15%. Compared to Table 1, those best power results are achieved using values for the maximum cluster size parameter greater than the true cluster size (see 'Population size in percentage (%)' in Table 1). Therefore, there is no evidence that values of the maximum cluster size parameter close to the real cluster size would improve power performance. This fact is also evident from the patterns of the dark gray cells. In general, choosing a maximum cluster size parameter of 50% generates power estimates very close to the best results.

A comparison of power for circular and elliptic scan statistics is shown in Figure 3. The horizontal axis represents all values of power for the circular scan statistic, whereas the vertical axis represents values of power for the elliptic scan statistic. The straight line represents the situation where the power of the elliptic and circular scan statistics is the same. Values above the straight line represent results where the elliptic scan statistic performs better than the circular scan statistic. Values below the straight line represents results where the circular scan statistic performs better than the elliptic scan statistic. It can be seen that the results for the irregular cluster models are above the straight lines, whereas circular results for the mixed, rural and urban cluster models are below the straight line, and close to one for both horizontal and vertical axes. Figure 3 shows that, for irregular cluster models, best results are achieved using the elliptic scan statistic, as expected. For circular cluster models, the circular scan statistic achieves better results, as expected. For the circular cluster models, power results are higher and close to one for both circular and elliptic scan statistics, although the circular scan statistic results are slightly better.

Sensitivity results are presented in Table 3. The results are very similar to the power results. It can be seen that, choosing a value of 50% for the maximum cluster size parameter, the results are very close to the best results. This is because of the detected cluster size. Having chosen a maximum

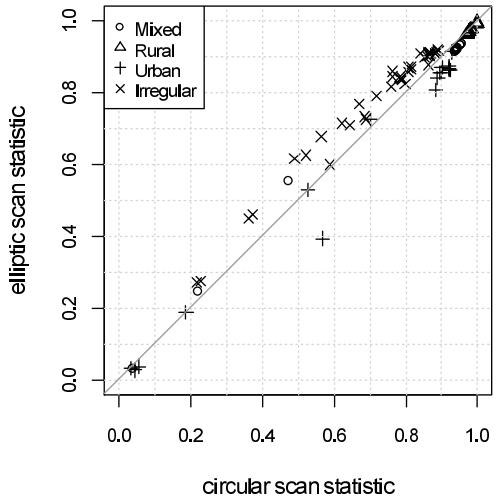
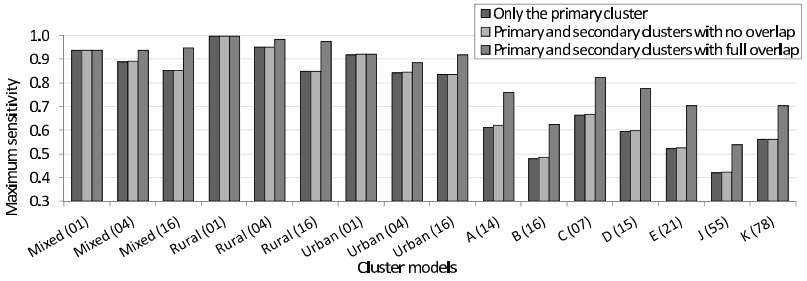


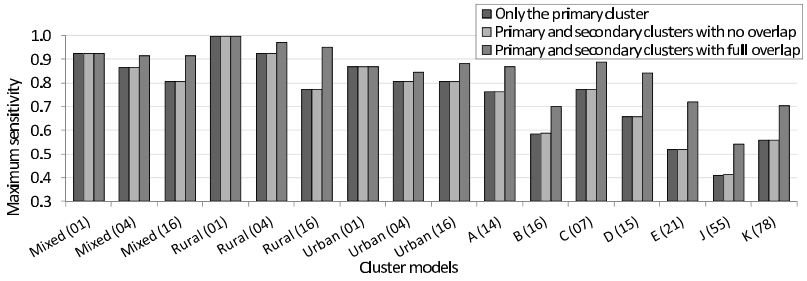
Figure 3: Comparison of the statistical power results for circular and elliptic scan statistics. Mixed, Rural and Urban represent the circular cluster models. The straight line represents the situation where the power of the elliptic and circular scan statistics is the same. In general, circular clusters are better detected using the circular scan statistic (results below the straight line), whereas irregular clusters are better detected using the elliptic scan statistic (results above the straight line).

cluster size of 50%, small values of PPV (see Table 4) and large values of misclassification (see Table 5) are generated. This means that the detected clusters are large. The chance that a large detected cluster contains elements of the true unknown cluster is high. Since the sensitivity statistic does not account for the detected cluster size, as long as the detected cluster size is large, values of sensitivity will be large too.

Figure 4 shows maximum values of sensitivity for both circular and elliptic scan statistics, accounting for secondary clusters in the cluster analysis. On the horizontal axis the circular and irregular cluster models are shown. It can be seen that there is a subtle difference between the results of the circular and elliptic scan statistics. As previously mentioned, the circular scan statistic provides better results for truly circular clusters. The same applies to the elliptic scan statistics if the true clusters are irregular. However, results show that if secondary clusters are included in the cluster analysis, the sensitivity results are improved, on average. These results are very intuitive: the more areas included in the detected cluster, the greater the chances that the areas of the true cluster are in the detected cluster. Nevertheless, accounting for



(a) circular scan statistic

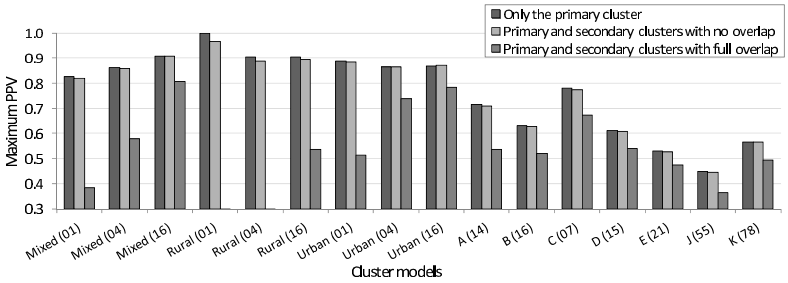


(b) elliptic scan statistic

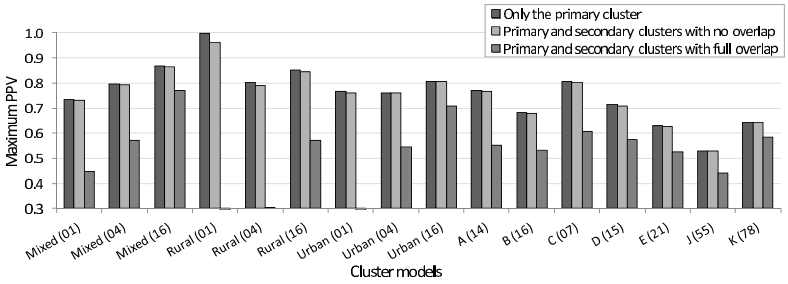
Figure 4: Best sensitivity results for circular and elliptic scan statistics, for different cluster models and accounting for the analysis of: primary cluster only; secondary clusters with no overlap with the primary cluster; and secondary clusters with no restrictions (full overlap)

secondary clusters with full overlap may generate detected clusters which are larger than the maximum cluster size parameter, i.e., if the maximum cluster size parameter is set as 50%, the detected cluster may be greater than this value.

PPV results are shown in Table 4. For the circular cluster models, the circular scan statistic most frequently achieves best results for maximum cluster size values smaller or equal to 5%. It can be argued that, for these scenarios, small values for the maximum cluster size parameter allow the method to detect small pieces of the true cluster. However, this pattern is not observed for the irregular cluster models nor for the elliptic scan statistic. Remarkably, in general, the double scan statistic achieves best PPV results. This is because, on average, the double scan statistic detects pieces of the true cluster better than either the circular scan statistic or the elliptic scan statistic. It is worth noting that for the irregular cluster models A and C the double scan statistic did not achieve good results. This is because these cluster models are very narrow and the counties of the clusters are usually



(a) circular scan statistic



(b) elliptic scan statistic

Figure 5: Best PPV results for circular and elliptic scan statistics, for different cluster models and accounting for the analysis of: primary cluster only; secondary clusters with no overlap with the primary cluster; and secondary clusters with no restrictions (full overlap)

not double connected.

Figure 5 shows best PPV results, accounting for the analysis of secondary clusters. Different from the sensitivity results, PPV values are not improved if secondary clusters are included into the cluster analysis. If the cluster analysis includes only secondary clusters with no overlap with the primary cluster, then there is a slight difference between the results. However, when primary and secondary clusters with full overlap are evaluated, then PPV results are even worse. These results suggest that if only one true cluster is present, secondary clusters, on average, will not detect pieces of the true cluster. Among all evaluated cluster models, the PPV statistic where slightly improved for the circular urban cluster models with size of 16 counties, and using the circular scan statistic.

Finally, misclassification results are shown in Table 5. Two distinct patterns are visualized. First, minimum values of misclassification are achieved for maximum cluster sizes of 2% and 5%. This is because of the misclassifica-

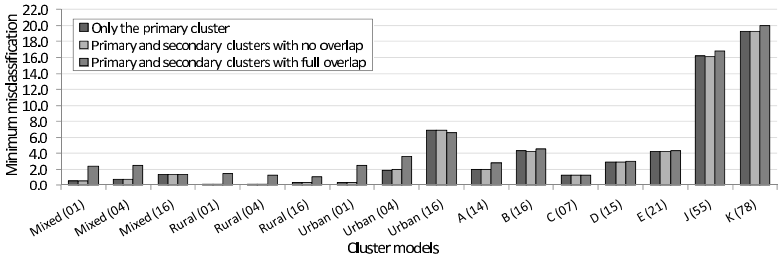
tion statistic which accounts for both the true cluster populations which were not detected, and the detected populations which do not belong to the true cluster. Recall that PPV results show that for small values of the maximum cluster size, circular and elliptic scan statistics do not perform well. Therefore, for large values of the maximum cluster size, the sum of the mistakenly detected or mistakenly missing populations is much larger than the sum of the mistakenly detected or mistakenly missing populations using smaller values of the maximum cluster size. Second, the double scan statistic achieves best results, or closer to the minimum values for both circular and irregular cluster models, on average.

Figure 6 shows that the more secondary clusters are included into the cluster analysis, the worse is the misclassification performance. This is because of the increase of the population in the detected cluster, if secondary clusters are included. Therefore, the misclassification statistics increases due to the increase of the detected populations which do not belong to the true cluster.

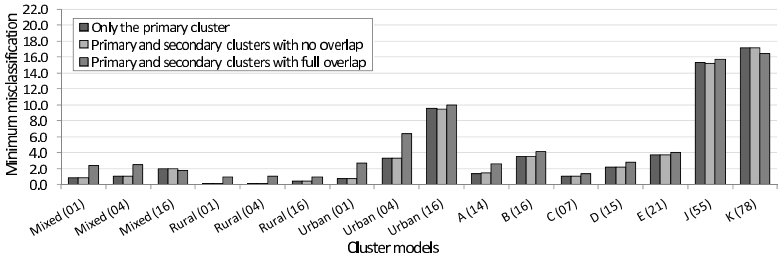
6. Discussion and Conclusion

The analysis of clusters through simulation studies is widely explored in the literature. In general, these simulation studies evaluate only primary clusters and generally apply a fixed value for the maximum cluster size. In this work, we compare the performance of different spatial scan statistics, using different values for the maximum cluster size, and considering different options for the analysis of secondary clusters. Results do suggest that performance measures are sensitive to the maximum cluster size chosen by the user.

Based on the simulation study, major findings are: first, there is evidence that the estimated power is, on average, insensitive to the maximum cluster parameter, although smaller values do compromise the estimated power. Therefore, empirical results suggest a maximum cluster size of 50%; Second, best sensitivity results are also achieve choosing a maximum cluster size of 50%, and evaluating significant secondary clusters with no restrictions regarding to overlapping. This is particularly tricky because the sensitivity statistic is generally improved when the detected cluster size increases. Therefore, as mentioned previously, the greater the detected cluster size the higher the chance to contain a piece of the true unknown cluster; Third, the PPV statistic measures the performance of the scan statistic to detect



(a) circular scan statistic



(b) elliptic scan statistic

Figure 6: Best misclassification results for circular and elliptic scan statistics, for different cluster models and accounting for the analysis of: primary cluster only; secondary clusters with no overlap with the primary cluster; and secondary clusters with no restrictions (full overlap)

the totality of the real cluster, and the results suggest that a small cluster size parameter must be chosen, such as 5% or, alternatively, the double scan statistic may be applied. In this case, secondary clusters may not be included into the cluster analysis. Finally, the misclassification results also suggest the double scan statistic with no analysis of secondary clusters.

Overall, there is evidence that the analysis of secondary clusters do not improve cluster detection performances, on average. Furthermore, there is no evidence that for truly irregular clusters the analysis of secondary clusters improves clustering performance.

We also apply different performance measures. In general, there is not a consensus in the literature about a unique performance measure. Therefore, the most appropriate performance measure is related to the goal of the cluster analysis, that is, before running the cluster analysis the user might be conscious about what he or she wants to find. Alternatively, it can be argued that the results provided in this work can be used to draw guidelines for practical use of scan statistics. That is, given a study region, the user may concern first about the power statistic. In this case, the circular scan statistic or the elliptical scan statistic can be applied with a maximum cluster size of 50%. If the detected cluster is significant and small then the cluster analysis is complete. Otherwise, if the detected cluster is significant and large then the user may be interested in detecting a small group of areas more likely to belong to the true cluster. To do so, the user may re-run the scan statistic or the elliptical scan statistic with a maximum cluster size of 5%, or run the double scan statistic.

Regarding the elliptic scan statistic, it has been shown in the literature that the elliptic scan statistic performs better for irregular cluster models than the circular scan statistic. Likewise, the circular scan statistics performs better for circular cluster models than scan statistics with irregular shape. In practice, the true cluster is unknown and the user has to decide whether using circular or elliptic scan statistics, or both.

It is worth noting that presented results are conditioned to the simulated data set that was used. Therefore, this work has limitations but also provides guidelines for simulation studies using different data sets. Simulation studies do give support to cluster analysis with real data sets. The underlying population and the total number of cases in real data sets can be used to design the simulation study and then provide empirical evidence about the optimal choice for the scan statistic parameters and shapes.

In conclusion, we believe that the presented simulation study provides

insights for epidemiologists to properly select among the circular, elliptic and double scan statistics, as well as to choose the parameter of the maximum size in order to improve cluster detection performance.

References

Assunção, R. M., Costa, M. A., Tavares, A., Ferreira, S., 2006. Fast detection of arbitrarily shaped disease clusters. Statistics in Medicine 25 (5), 723–742.

Carpenter, T. E., 2011. The spatial epidemiology (r)evolution: A look back in time and forward to the future. Spatial and Spatio-temporal Epidemiology 2, 119–124.

Chaput, E. K., Meek, J. I., Heimer, R., 2002. Spatial analysis of Human Granulocytic Ehrlichiosis near Lyme, Connecticut. Emerging Infectious Diseases 8 (9), 943–948.

Costa, M. A., Assunção, R. M., 2005. A fair comparison between the spatial scan and the Besag-Newell disease clustering tests. Environmental and Ecological Statistics 12 (3), 301–319.

Costa, M. A., Assunção, R. M., Kulldorff, M., 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Computational Statistics and Data Analysis 56, 1771–1783.

Costa, M. A., Kulldorff, M., 2009. Scan Statistics: Methods and Applications. Statistics for industry and technology. Birkkäuser, Ch. 6, pp. 129–152.

Coulston, J. W., Riitters, K. H., 2003. Geographic analysis of forest health indicators using spatial scan statistics. Environmental Management 31 (6), 764–773.

Donnan, P. T., Parratt, J. D. E., Wilson, S. V., Forbes, R. B., O'Riordan, J. I., Swingler, R. J., 2005. Multiple sclerosis in Tayside, Scotland: detection of clusters using a spatial scan statistic. Multiple Sclerosis Journal 11, 403–408.

Duczmal, L., Assunção, R. M., 2004. Simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics and Data Analysis 45 (2), 269–286.

Duczmal, L., Cançado, A. L., Takahashi, R. H., Bessegato, L. F., 2007. A genetic algorithm for irregularly shaped spatial scan statistics. Computational Statistics and Data Analysis 52 (1), 43–52.

Duczmal, L., Kulldorff, M., Huang, L., 2006. Evaluation of spatial scan statistics for irregularly shaped disease clusters. Journal of Computational and Graphical Statistics 15, 428–442.

Dwass, M., 1957. Modified randomization tests for nonparametric hypothesis. Annals of Mathematical Statistics 28, 181–187.

Forand, S. P., Talbot, T. O., Druschel, C., Cross, P. K., 2002. Data quality and the spatial analysis of disease rates: congenital malformations in New York State. Health and Place 8, 191–199.

Huang, L., Kulldorff, M., Gregorio, D., 2007. A spatial scan statistic for survival data. Biometrics 63 (1), 109–118.

Jung, I., Kulldor, M., Richard, O. J., 2010. A spatial scan statistic for multinomial data. Statistics in Medicine 29 (18), 1910–1918.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics: Theory and Methods 26 (6), 1481–1496.

Kulldorff, M., July 2010. SaTScan User Guide for version 9.0.

Kulldorff, M., Nagarwalla, N., 1995. Spatial disease clusters: Detection and inference. Statistics in Medicine 14, 799–810.

Kulldorff, M., Pickle, L., Huang, L., Duczmal, L., 2006. An elliptic spatial scan statistic. Statistics in Medicine 25, 3929–3943.

Kulldorff, M., Tango, T., Park, P. J., 2003. Power comparisons for disease clustering tests. Computational Statistics and Data Analysis 42, 665–68.

Kulldorff, M., Zhang, Z., Hartman, J., Heffernan, R., Huang, L., Mostashari, F., 2004. Benchmark data and power calculations for evaluating disease outbreak detection methods. Morbidity and Mortality Weekly Report 53, 144–151.

Naus, J. I., 1965. Clustering of random points in two dimensions. Biometrika 52, 263–267.

Patil, G. P., Taillie, C., 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics 11, 183–197.

Que, J., Tsui, F. C., Espino, J., 2008. A Z-score based multi-level spatial clustering algorithm for the detection of disease outbreaks. Lecture Notes in Computer Science 5354, 108–118.

Read, S., Bath, P., Willet, P., Maheswaran, R., 2011. Measuring the spatial accuracy of the spatial scan statistic. Spatial and Spatio-temporal Epidemiology 2, 69–78.

Song, C., Kulldorff, M., 2003. Power evaluation of disease clustering test. International Journal of Health Geographic 2 (9).

Takahashi, K., Kulldorff, M., Tango, T., Yih, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics 4 (11).

Takahashi, K., Kulldorff, M., Tango, T., Yih, K., 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. International Journal of Health Geographics 7 (14).

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L., Clark, L. C., 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. American Journal of Epidemiology 132 (1), 136–143.

Table 1: Simulated cluster models.

Type	Region	Number of counties	Population size	Population size in percentage (%)	Relative risk
Circular clusters	Mixed	1	710,196	2.40%	2.85
		4	1,108,440	3.75%	2.4
		16	1,684,327	5.70%	2.1
	Rural	1	2,675	0.01%	192.89
		4	132,343	0.45%	7.05
		16	360,275	1.22%	3.90
	Urban	1	786,178	2.66%	2.73
		4	2,953,077	10.00%	1.81
		16	7,627,173	25.82%	1.53
Irregular clusters	A	14	1,057,407	3.58%	2.32
	B	16	1,672,387	5.66%	1.97
	C	7	709,519	2.40%	2.71
	D	15	1,119,235	3.79%	2.29
	E	21	1,483,995	5.02%	2.06
	J	55	3,198,049	10.83%	1.63
	k	78	7,775,129	26.32%	1.34

Table 2: Statistical power for circular, elliptic and double scan statistics. Boldface values indicate results with highest power for each row. The cells whose values are no greater than of 0.01 (1%) distance from the boldface value are shaded in dark gray. The cells whose values are within 0.01 and 0.02 (1% – 2%) distance from the boldface value, are shaded in light gray.

Cluster model		Maximum cluster size						Double
		2%	5%	10%	15%	25%	50%	
Circular scan statistic								
Circular clusters	#counties							
Mixed	1	0.0364	**0.9404**	0.9392	0.9386	0.9383	0.9357	0.9154
	4	0.2205	0.9429	**0.9445**	0.9394	0.9389	0.9375	0.9293
	16	0.4730	0.9340	**0.9539**	0.9531	0.9518	0.9492	0.9166
Rural	1	0.9981	0.9981	0.9982	0.9982	0.9983	**0.9984**	**0.9984**
	4	**0.9849**	0.9777	0.9745	0.9744	0.9736	0.9725	0.9694
	16	**0.9802**	0.9778	0.9746	0.9730	0.9717	0.9695	0.9197
Urban	1	0.0340	0.9189	0.9219	0.9204	0.9210	**0.9226**	0.9004
	4	0.0454	0.5680	0.8820	0.8889	**0.8976**	0.8947	0.8487
	16	0.0569	0.1872	0.5271	0.7017	0.9004	**0.9266**	0.8120
Irregularly shaped clusters								
A	14	0.6856	0.8636	**0.8688**	0.8682	0.8676	0.8530	0.8146
B	16	0.5878	0.7583	0.7828	0.7893	0.7859	0.7878	**0.8002**
C	7	0.8641	0.8879	**0.8908**	0.8863	0.8808	0.8808	0.7868
D	15	0.7627	0.8418	0.8646	0.8671	0.8639	0.8605	**0.8922**
E	21	0.6713	0.7620	0.8082	**0.8146**	0.8139	0.8068	0.7770
J	55	0.2271	0.3621	0.4901	0.5642	0.6428	**0.6875**	0.5940
K	78	0.2186	0.3731	0.5204	0.6205	0.7172	**0.7978**	0.5259
Elliptic scan statistic								
Circular clusters	#counties							
Mixed	1	0.0314	**0.9275**	0.9258	0.9232	0.9191	0.9164	0.9154
	4	0.2463	0.9239	**0.9282**	0.9258	0.9228	0.9181	0.9293
	16	0.5558	0.9199	**0.9394**	0.9378	0.9355	0.9337	0.9166
Rural	1	0.9981	0.9922	0.9923	0.9923	0.9923	0.9923	**0.9984**
	4	**0.9768**	0.9686	0.9656	0.9647	0.9637	0.9623	0.9694
	16	**0.9780**	0.9717	0.9677	0.9664	0.9644	0.9626	0.9197
Urban	1	0.0319	0.8737	0.8687	0.8666	0.8649	0.8647	**0.9004**
	4	0.0267	0.3937	0.8067	0.8416	0.8557	**0.8572**	0.8487
	16	0.0375	0.1864	0.5305	0.7265	0.8706	**0.9136**	0.8120
Irregularly shaped clusters								
A	14	0.7333	**0.9080**	0.9098	0.9065	0.9026	0.8998	0.8146
B	16	0.5996	0.8179	**0.8430**	0.8414	0.8392	0.8360	0.8002
C	7	0.8780	**0.9214**	0.9174	0.9136	0.9106	0.9087	0.7868
D	15	0.8447	0.9097	**0.9127**	0.9125	0.9080	0.9024	0.8922
E	21	0.7690	0.8599	0.8707	**0.8720**	0.8647	0.8583	0.7770
J	55	0.2736	0.4491	0.6169	0.6776	0.7108	**0.7260**	0.5940
K	78	0.2700	0.4616	0.6259	0.7151	0.7908	**0.8242**	0.5259

Table 3: Sensitivity results for circular, elliptic and double scan statistics. Boldface values indicate results with the highest sensitivity for each row. The cells whose values are no greater than of 0.01 (1%) distance from the boldface value are shaded in dark gray. The cells whose values are within 0.01 and 0.02 (1% – 2%) distance from the boldface value, are shaded in light gray.

Cluster model		Maximum cluster size						double
		2%	5%	10%	15%	25%	50%	
Circular scan statistic								
Circular clusters	#counties							
Mixed	1	0.0000	**0.9369**	0.9363	0.9358	0.9358	0.9335	0.9115
	4	0.0364	0.8836	**0.8890**	0.8847	0.8847	0.8835	0.8657
	16	0.0885	0.7151	0.8513	**0.8526**	0.8523	0.8504	0.7103
Rural	1	**0.9980**	**0.9980**	**0.9980**	**0.9980**	**0.9980**	**0.9980**	0.9979
	4	**0.9506**	0.9435	0.9409	0.9405	0.9398	0.9392	0.9019
	16	0.8477	**0.8484**	0.8466	0.8453	0.8445	0.8430	0.5656
Urban	1	0.0000	0.9147	0.9183	0.9166	0.9174	**0.9194**	0.8963
	4	0.0019	0.2391	0.8202	0.8281	**0.8433**	0.8424	0.7473
	16	0.0019	0.0247	0.1638	0.3280	0.7570	**0.8353**	0.5582
Irregularly shaped clusters								
A	14	0.2299	0.5810	0.6013	0.6074	**0.6118**	0.6044	0.5018
B	16	0.1282	0.3122	0.4338	0.4602	0.4678	**0.4783**	0.4274
C	7	0.5427	0.6384	**0.6630**	0.6617	0.6594	0.6610	0.4781
D	15	0.2333	0.4607	0.5802	0.5904	0.5941	**0.5947**	0.5487
E	21	0.1511	0.3168	0.4842	0.5088	0.5196	**0.5206**	0.3291
J	55	0.0114	0.0539	0.1282	0.2102	0.3166	**0.4194**	0.1907
K	78	0.0088	0.0417	0.1110	0.2008	0.3411	**0.5605**	0.1168
Elliptic scan statistic								
Circular clusters	#counties							
Mixed	1	0.0000	**0.9242**	0.9230	0.9209	0.9173	0.9147	0.9115
	4	0.0511	0.8526	**0.8662**	0.8645	0.8621	0.8581	0.8657
	16	0.1285	0.6850	0.8021	0.8051	**0.8052**	0.8041	0.7103
Rural	1	**0.9980**	0.9921	0.9922	0.9922	0.9922	0.9922	0.9979
	4	**0.9236**	0.9165	0.9135	0.9130	0.9122	0.9112	0.9019
	16	0.7712	**0.7736**	0.7713	0.7709	0.7693	0.7681	0.5656
Urban	1	0.0000	0.8672	0.8624	0.8608	0.8599	0.8601	**0.8963**
	4	0.0005	0.1608	0.7047	0.7749	0.7997	**0.8046**	0.7473
	16	0.0013	0.0279	0.1711	0.3558	0.6851	**0.8043**	0.5582
Irregularly shaped clusters								
A	14	0.2948	0.7504	**0.7615**	0.7602	0.7580	0.7565	0.5018
B	16	0.1423	0.4731	0.5733	0.5786	0.5830	**0.5854**	0.4274
C	7	0.6021	0.7689	**0.7710**	0.7688	0.7674	0.7663	0.4781
D	15	0.3129	0.6167	0.6519	**0.6566**	0.6562	0.6543	0.5487
E	21	0.2031	0.4238	0.4943	0.5119	**0.5174**	0.5172	0.3291
J	55	0.0165	0.0717	0.1917	0.2771	0.3572	**0.4113**	0.1907
K	78	0.0130	0.0569	0.1492	0.2533	0.4386	**0.5576**	0.1168

Table 4: PPV results for circular, elliptic and double scan statistics. Boldface values indicate results with the highest PPV values for each row. The cells whose values are no greater than of 0.01 (1%) distance from the boldface value are shaded in dark gray. The cells whose values are within 0.01 and 0.02 (1% – 2%) distance from the boldface value, are shaded in light gray.

Cluster model		Maximum cluster size						double
		2%	5%	10%	15%	25%	50%	
Circular scan statistic								
Circular clusters	#counties							
Mixed	1	0.0000	**0.8250**	0.8131	0.8112	0.8097	0.8072	0.8167
	4	0.1472	**0.8613**	0.8323	0.8240	0.8211	0.8183	0.8514
	16	0.4113	**0.9092**	0.8791	0.8701	0.8650	0.8606	0.8904
Rural	1	**0.9979**	**0.9979**	**0.9979**	**0.9979**	**0.9979**	**0.9979**	0.9978
	4	0.9030	0.8939	0.8911	0.8908	0.8902	0.8891	**0.9110**
	16	**0.9035**	0.8882	0.8831	0.8812	0.8794	0.8773	0.8343
Urban	1	0.0000	**0.8882**	0.8738	0.8689	0.8660	0.8638	0.8356
	4	0.0200	0.5555	**0.8648**	0.8426	0.8153	0.7955	0.6911
	16	0.0367	0.1709	0.5058	0.6693	**0.8698**	0.8302	0.7853
Irregularly shaped clusters								
A	14	0.6183	**0.7171**	0.6800	0.6666	0.6570	0.6413	0.6401
B	16	0.5057	0.6309	0.5811	0.5588	0.5423	0.5319	**0.6717**
C	7	**0.7812**	0.7130	0.6835	0.6746	0.6668	0.6639	0.6912
D	15	0.6116	0.5884	0.5402	0.5290	0.5186	0.5124	**0.7466**
E	21	0.5292	0.5213	0.4834	0.4681	0.4554	0.4460	**0.6257**
J	55	0.1895	0.3079	0.3927	0.4267	0.4473	0.4288	**0.5041**
K	78	0.1951	0.3342	0.4438	0.5179	**0.5669**	0.5608	0.4669
Elliptic scan statistic								
Circular clusters	#counties							
Mixed	1	0.0000	0.7343	0.7080	0.7021	0.6975	0.6942	**0.8167**
	4	0.1642	0.7957	0.7402	0.7310	0.7256	0.7205	**0.8514**
	16	0.4834	0.8677	0.8024	0.7869	0.7792	0.7753	**0.8904**
Rural	1	0.9974	0.9915	0.9915	0.9915	0.9915	0.9915	**0.9978**
	4	0.8025	0.7931	0.7899	0.7892	0.7885	0.7878	**0.9110**
	16	**0.8523**	0.8267	0.8209	0.8191	0.8172	0.8153	0.8343
Urban	1	0.0000	0.7670	0.7227	0.7123	0.7064	0.7013	**0.8356**
	4	0.0045	0.3651	**0.7595**	0.7161	0.6774	0.6595	0.6911
	16	0.0207	0.1681	0.5051	0.6915	**0.8064**	0.7567	0.7853
Irregularly shaped clusters								
A	14	0.6417	**0.7708**	0.7320	0.7226	0.7162	0.7120	0.6401
B	16	0.4998	**0.6812**	0.6372	0.6187	0.6067	0.5973	0.6717
C	7	**0.8074**	0.7456	0.7195	0.7127	0.7085	0.7058	0.6912
D	15	0.7151	0.7059	0.6550	0.6450	0.6362	0.6303	**0.7466**
E	21	**0.6287**	**0.6287**	0.5677	0.5501	0.5358	0.5287	0.6257
J	55	0.2246	0.3780	0.5060	**0.5283**	0.5110	0.4928	0.5041
K	78	0.2372	0.4100	0.5393	0.6033	**0.6438**	0.6288	0.4669

Table 5: Misclassification results for circular, elliptic and double scan statistics. Boldface values indicate results with the lowest misclassification results for each line. The cells whose values are no greater than of 0.01 (1%) distance from the boldface value are shaded in dark gray. The cells whose values are within 0.01 and 0.02 (1% – 2%) distance from the boldface value, are shaded in light gray.

Cluster model		Maximum cluster size						double
		2%	5%	10%	15%	25%	50%	
Circular scan statistic								
Circular clusters	#counties							
Mixed	1	2.4289	0.5599	0.6702	0.7053	0.7506	0.8022	**0.5368**
	4	3.7019	**0.7770**	1.0026	1.0746	1.1492	1.2344	0.8292
	16	5.2755	1.7200	**1.3589**	1.4831	1.5883	1.6919	1.7669
Rural	1	**0.0002**	**0.0002**	0.0010	0.0010	0.0028	0.0075	0.0027
	4	**0.0787**	0.0918	0.0988	0.1069	0.1137	0.1263	0.0882
	16	**0.2941**	0.3396	0.3670	0.3793	0.3934	0.4088	0.6444
Urban	1	2.6845	**0.3406**	0.4881	0.5529	0.6447	0.8362	0.6665
	4	9.9982	7.6259	**1.9200**	2.2687	2.9055	3.7327	4.6231
	16	25.7861	25.2098	21.6991	17.6919	**6.8770**	7.4745	11.8296
Irregularly shaped clusters								
A	14	2.8384	**2.0149**	2.3562	2.5974	2.8893	3.1146	2.5378
B	16	5.0516	4.3054	4.3728	4.8073	5.2454	6.0270	**3.9263**
C	7	**1.2356**	1.4143	1.6948	1.8063	1.9339	2.1132	1.4734
D	15	3.1286	2.8992	3.3469	3.6402	3.9819	4.2657	**2.2701**
E	21	4.4774	4.2709	4.6417	5.0490	5.5493	5.9663	**3.9152**
J	55	18.7500	18.0507	17.1200	16.3677	16.1608	17.9380	**15.9780**
K	78	26.1171	25.3493	23.9545	22.1722	20.0801	**19.3198**	23.6712
Elliptic scan statistic								
Circular clusters	#counties							
Mixed	1	2.4339	0.8567	1.1157	1.2020	1.2778	1.3927	**0.5368**
	4	3.6653	1.0995	1.5740	1.7211	1.8414	1.9682	**0.8292**
	16	5.0752	2.0126	2.0930	2.3146	2.4931	2.6290	**1.7669**
Rural	1	**0.0005**	0.0005	0.0016	0.0016	0.0016	0.0016	0.0027
	4	0.1687	0.1889	0.2068	0.2126	0.2237	0.2316	**0.0882**
	16	**0.4554**	0.5351	0.5696	0.5926	0.6071	0.6349	0.6444
Urban	1	2.6938	0.7773	1.1508	1.3390	1.5452	1.9047	**0.6665**
	4	10.0147	8.4792	**3.3235**	3.7587	4.8421	5.8690	4.6231
	16	25.8035	25.1526	21.5832	17.0508	**9.5175**	10.3266	11.8296
Irregularly shaped clusters								
A	14	2.6667	**1.4296**	1.7813	1.9205	2.0547	2.2120	2.5378
B	16	5.0086	**3.5139**	3.6970	4.0224	4.3894	4.8844	3.9263
C	7	**1.0717**	1.1269	1.3600	1.4490	1.5286	1.6318	1.4734
D	15	2.8136	**2.2086**	2.6801	2.8942	3.0989	3.2981	2.2701
E	21	4.2279	**3.7589**	4.3066	4.6906	5.0834	5.3575	3.9152
J	55	18.6789	17.8018	16.0660	**15.2659**	15.4534	16.5002	15.9780
K	78	26.0271	25.0076	23.0448	20.9364	17.5590	**17.1875**	23.6712