An Alternative Cluster Detection Test in Spatial Scan Statistics

Ahmed Reza Soltani* Suja Mansour Aboukhamseen†

Abstract

We establish a hypotheses testing procedure equivalent to the Kulldorff (1997) spatial scan hypotheses test for cluster detection, then provide transparent test statistics for cluster detection in a spatial setting. We also specify the limiting distribution of the test statistics. We apply our method to North Carolina sudden infant death syndrome cases; it detects the same primary and secondary clusters as Kulldorff (1997). Simulated data is used to compare the performance of our method with that of Kulldorff and the findings show that our test is more sensitive and accurate in detecting clusters.

Key Words: Spatial scan statistics, Cluster detection, Test statistics, Limiting Distribution

1. Introduction

The problem which is of interest and of practical importance in spatial scan statistics is to test whether a *region* Z of *individuals* is a cluster of a certain characteristic, called a *point*. This issue is discussed by several authors including Naus (1965), Glaz and Naus (1983), Loader (1991), Tango and Takahashi (2005), and Zhang and Lin (2009). Kulldorff (1997) provides a cluster finding procedure when there are only two possibilities for an individual in a region Z, and in a predetermined region outside Z; an individual is either a point or is not a point. An individual in Z is a point with probability p, while in Z^c, the complement of Z, an individual is a point with probability q. Kulldorff (1997) considers the test $H_0 : \ p = q$ *vs* $H_1 : \ p > q$. A parameter is a triple (Z, p, q); the parameter space is $\mathcal{S} = \{(Z_i, p, q); 0 < p < 1, \ 0 < q < 1, \ p \geq q, \ i = 1, ..., J\}$ which is partitioned by the hypothesis into $\mathcal{S}_0 = \{(Z_i, p, q); p = q, i = 1, ..., J\}$ and $\mathcal{S}_1 = \{(Z_i, p, q); p > q, i = 1, ..., J\}$. According to Kulldorff (1997) a region $\hat{Z}$ among $Z_1, ..., Z_J$ is a cluster if it maximizes $L(Z) = \sup_{p>q} L[Z, p, q]$ where $L[Z, p, q] \equiv L[Z, p, q : n(Z), n(G)]$ is the

*Department of Statistics and Operations Research, Faculty of Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait.
This research was supported by the Kuwait University Research Adminstration under research grant SS01/10.
†Department of Statistics and Operations Research, Faculty of Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait email: suja.aboukhamseen@ku.edu.kw

likelihood function. The observed sample is $[n(Z_i), n(G)],\ i = 1, ..., J$, where $n(Z_i)$ and $n(G)$ are the total number of observed individuals in Z_i and G respectively. For the statistical inference, Kulldorff (1997) uses the ratio $\lambda = L(\hat{Z})/L_0$, where $L_0 = \sup_{p=q} L[Z, p, q]$. Conceivably, L_0 does not depend on Z, as for $p = q$ it is only $n(G)$ that matters. Simulated values of λ are used to obtain critical values of the test statistic.

In our understanding of the Kulldorff (1997) basic settings of the spacial scan statistic, the parameters p and q indeed depend on the region Z. In other words, Z is an independent parameter but p and q are dependent parameters. Therefore we interpret p, or q, as the conditional probability that an individual is a point given that the individual is in Z, or in Z^c; and denote them by $p \equiv P(+|Z)$ and $q \equiv P(+|Z^c)$, respectively. By using this setting we read the null and its alternative hypotheses as

$$\mathcal{A}: \quad H_0 : P(+|Z) = P(+|Z^c) \quad \text{vs} \quad H_1 : P(+|Z) > P(+|Z^c). \tag{1.1}$$

Under H_0 the region Z is not contaminated or is not a cluster, since knowing that an individual is in Z does not increase the probability that the individual is a point. *We call a region a cluster if H_0 is rejected in favor of H_1, in (1.1).*

In this article we provide transparent procedures for testing hypothesis $\mathcal{A}$; and ultimately cluster finding using spatial scan statistics. In our procedure, the test statistics and their distributions are fully specified. For a small sample, the Fisher Exact Test may be simply applied. For a large sample, an adjusted test statistic for the population proportion with a limiting normal distribution provide effective results. In contrast to the Kulldorff (1997) where the distribution of the corresponding test statistic is not specified and the critical values are computed empirically through simulation, we provide the exact and asymptotic distribution of our test statistics. In order to check the effectiveness of our procedure, we apply it to simulated data as well as the sudden infant death syndrome (SIDS) cases, in Cressie and Chan (1989). Our procedure detects exactly the same primary and secondary cluster regions as Kulldorff (1997). We also provide the power of the test statistic.

This article is organized as follows. In Section 2 we establish our formulation for the hypotheses on a suspected spatial cluster. In Section 3 we present our clustering method and provide the corresponding test statistics. In Section 4 we present the asymptotic distribution of the test statistics; we devote Section 5 to numerical derivations and applications using the open-source statistical computation environment R (R Development Core Team, 2010). We conclude the article with a discussion.

2. The New Formulation

Assume that G is a region. Each individual in G carries one and only one of K possible labels; thus individuals are classified into K classes. The k^{th} class consists of individuals that carry label k, $k = 1, \ldots, K$. For a subregion Z in G, we let $n_k(Z)$ denote the total number of individuals with label k in Z, and B_k denote the event that a label is of type k, $k = 1, \ldots, K$. Also, we let A_Z be the event that an individual is in a region Z. We define $P_{k|Z} \equiv P(B_k|A_Z)$ as the conditional probability that an individual is labeled k, $k = 1, \ldots, K$, given that the individual is in Z. We assume the total number of individuals in Z, denoted by $\mu(Z)$, is known. We rephrase the hypothesis $\mathcal{A}$ in (1.1) as

$$H_0 : P_{k|Z} = P_{k|Z^c} \quad \text{vs} \quad H_1 : P_{k|Z} > P_{k|Z^c}.$$

This is exactly the same as the hypotheses described in (1.1) in the Introduction, an individual is a point if it has label k.

3. A New Cluster Detection Method

In this section we present our method for cluster detection. We assume that the number of individuals in G, $\mu(G)$, and the number of individuals in Z, $\mu(Z)$, are known. We use the symbol $+$ to signify a label of interest, among K possible labels. An individual with a label $+$ is defined as a point. We use notations introduced in previous sections. The probability of the event A_Z, denoted by $\nu(Z)$, is clearly equal to $\nu(Z) = \frac{\mu(Z)}{\mu(G)}$.

Our method is based on the following crucial lemma. Indeed it allows us to express the hypotheses on $P_{+|Z}$ equivalently by hypotheses on $P_{Z|+}$.

Lemma 1. For a region Z and a label $+$, among K possible labels,

$$P_{+|Z} = P_{+|Z^c}, \tag{3.1}$$

if and only if

$$P_{Z|+} = \nu(Z). \tag{3.2}$$

Proof. Let us assume (3.1). We note that $P_{+|Z} = \frac{P_{Z|+}P(+)}{\nu(Z)}$, and $P_{+|Z^c} = \frac{P_{Z^c|+}P(+)}{\nu(Z^c)}$. Therefore,

$$\frac{P_{Z|+}P(+)}{\nu(Z)} = \frac{P_{Z^c|+}P(+)}{\nu(Z^c)}.$$

Equivalently,

$$\frac{P_{Z|+}}{\nu(Z)} = \frac{P_{Z^c|+}}{\nu(Z^c)} \Rightarrow \frac{P_{Z|+}}{\nu(Z)} = \frac{1}{\nu(Z^c)} - \frac{P_{Z|+}}{\nu(Z^c)}.$$

Therefore

$$P_{Z|+}\left[\frac{1}{\nu(Z)} + \frac{1}{\nu(Z^c)}\right] = \frac{1}{\nu(Z^c)}.$$

This leads us to

$$P_{Z|+} = \frac{\frac{1}{\nu(Z^c)}}{\left[\frac{1}{\nu(Z)} + \frac{1}{\nu(Z^c)}\right]} = \nu(Z),$$

giving (3.2). By reversing the argument given above, we easily derive (3.1) from (3.2). The proof is complete. □

Remark 1. If there are only two labels $+$ and $-$ then (3.2) is indeed equivalent to $P_{Z|-} = \nu(Z)$. This can be seen as follows. If (3.1) is satisfied then $P_{-|Z} = P_{-|Z^c}$. Thus by applying Lemma 3.1 to the label $-$, it follows that $P_{Z|-} = \nu(Z)$.

Lemma 2. For a region Z and a label $+$, among K possible labels,

$$P_{+|Z} > P_{+|Z^c}, \tag{3.3}$$

if and only if

$$P_{Z|+} > \nu(Z), \tag{3.4}$$

.

Proof. Let us assume (3.3). We note that $P_{+|Z} = \frac{P_{Z|+}P(+)}{\nu(Z)}$, and $P_{+|Z^c} = \frac{P_{Z^c|+}P(+)}{\nu(Z^c)}$. Therefore it follows from (3.3) that

$$\frac{P_{Z|+}P(+)}{\nu(Z)} > \frac{P_{Z^c|+}P(+)}{\nu(Z^c)}.$$

Equivalently,

$$\frac{P_{Z|+}}{\nu(Z)} > \frac{P_{Z^c|+}}{\nu(Z^c)} \Rightarrow \frac{P_{Z|+}}{\nu(Z)} = \frac{1}{\nu(Z^c)} - \frac{P_{Z|+}}{\nu(Z^c)}.$$

Therefore

$$P_{Z|+}\left[\frac{1}{\nu(Z)} + \frac{1}{\nu(Z^c)}\right] > \frac{1}{\nu(Z^c)}.$$

This leads us to

$$P_{Z|+} > \frac{\frac{1}{\nu(Z^c)}}{\left[\frac{1}{\nu(Z)} + \frac{1}{\nu(Z^c)}\right]} = \nu(Z),$$

giving (3.4). By reversing the argument given above, we easily derive (3.3) from (3.4). The proof is complete. □

Lemma 1 and Lemma 2 allow us to restate hypothesis $\mathcal{A}$ as:

$$\mathcal{B}: \quad H_0 : P_{Z|+} = \nu(Z) \quad \text{versus} \quad H_1 : P_{Z|+} > \nu(Z).$$

A zone Z is a cluster if, in $\mathcal{B}$, the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1.

Lemma 1 has an immediate significant implication, so we state it as a Lemma below.

Lemma 3. For a region Z and a label $+$, among possible K labels, (3.1) is equivalent to

$$P_{+,Z} = P_+ P_Z, \tag{3.5}$$

where $P_{+,Z}$ is the joint probability that an individual is labeled $+$ and is in Z; P_+ and P_Z are corresponding marginal probabilities.

Proof. Assume (3.1) is satisfied, then it follows from Lemma 1 that (3.2) is satisfied. Therefore $P_{+,Z} = \nu(Z)P_+$. But note that $\nu(Z) = P_Z$. Thus we arrive at (3.5). By reversing this reasoning we easily derive (3.1) from (3.5). The proof is complete. □

Remark 2. If there are only two labels $+$ and $-$, then it follows from Remark 1 and Lemma 3 that (3.1) is satisfied if and only if the label categories $\{+, -\}$ and region categories $\{Z, Z^c\}$ are independent.

4. Exact and Limiting Distributions

In the following lemma we provide, the exact and limiting distribution of number of individuals in a region Z that carry label $+$ under H_0 in Hypotheses $\mathcal{B}$.

Lemma 4.1. Assuming that $n_+(G)$, the total number of individuals carrying label $+$, is known, under H_0 in the hypotheses $\mathcal{B}$ we have the following;

(a): the exact distribution of $X_{Z,+}$, the number of individuals in region Z that carry label $+$, is a binomial distribution with parameters $n_+(G)$ and $\nu(Z)$; $X_{Z,+} \sim B(n_+(G), \nu(Z))$;

(b): the limiting distribution of $X_{Z,+}$ is normal with mean $n_+(G)\nu(Z)$ and standard deviation $\sqrt{n_+(G)\nu(Z)[1-\nu(Z)]}$.

Proof. The proof is an immediate consequence of the following lemma concerning the conditional probability in the multinomial distribution, and the null hypothesis in $\mathcal{B}$.

Lemma 4.2. In a multinomial distribution if r, the total of points in certain cells $Y_1, ..., Y_u$ is given, then the number of points in the cell Y_j follows a binomial distribution with mean $p_{Y_j} / \sum_{i=1}^{u} p_{Y_i}$.

According to Lemma 3 and Lemma 4.1, the following test procedures can be effectively applied to test hypotheses $\mathcal{B}$, and consequently be used to detect clusters; (i): the Fisher exact test; (ii): the binomial exact test for small to moderate sample size; (iii): the test on population proportion using the asymptotic normal distribution for large sample size. We use (iii). Thus according to Lemma 4.1 (b) a region Z is a cluster at level of significant α if $z = [X_{Z,+}/n_+(G) - \nu(Z)]/\sqrt{\nu(Z)[1-\nu(Z)]/n_+(G)}$ exceeds z_α.

The algorithm given below describes how the test statistic would be used in application:

1. Define the spatial region G under investigation.
2. Count the total population in G, $\mu(G)$, and the number of points of interest, $n_+(G)$.
3. For a given Z in G count the population in Z, $\mu(Z)$, and the number of points in Z, $X_{Z,+} = n_+(Z)$.
4. Compute the test statistic for hypothesis $\mathcal{B}$;
$$z = \frac{X_{Z,+}/n_+(G) - \nu(Z)}{\sqrt{\nu(Z)[1-\nu(Z)]/n_+(G)}}$$
5. Reject H_0 in $\mathcal{B}$ (or $\mathcal{A}$) if the p-value is greater than some predefined significance level.
6. Repeat 3 to 5 for all possible regions Z.

We apply our cluster detection method to both real and simulated data. We provide details in Section 5.

5. Application and Numerical Computations

In application, our procedure remains true to the nature of spatial scan statistics. A moving scan window, Z, is defined on a spatial region, G. The process of scanning the entire geographic region G may become

computationally intensive since the number of possible windows, Z, may reach infinity. The process is specified in part by the geometry of the area being scanned, and also by the size and shape of the scanning window. Both aspects are determined by the application at hand.

For the tests proposed in this paper, the window may take any predefined shape and vary in size. Once the test statistic is computed for *all possible windows*, then the cluster region is the region defined by the window with the highest significance.

For comparative purposes, we adhere to the methods used by Kulldorff and Nagarwalla (1995), and Kulldorff (1997) to scan a region. The scanning procedure they proposed was inspired largely by Openshaw's Geographic Analysis Machine (GAM) (Openshaw, Charlton, Wymer, & Craft, 1987), and a generalization of Turnbull's method (Turnbull, Iwano, Burnett, Howe, & Clark, 1990). The scanning window is, therefore, defined as a circle with an upper limit imposed on the size of the circle. By using this scanning technique, the number of scan windows is markedly curtailed.

Their process entails defining a series of foci on the coordinate grid of a geographic region G. A scan window is defined as a moving circle of varying size with a foci as its centroid. For a given center an increase in radius is only of interest if a new point enters the circle. The size of the circle is limited such that its coverage area does not encompass more than R, where R is a fraction of the total population in the region G.

The algorithm is described below:

1. A foci is selected arbitrarily to be the center of the scanning window. Calculate the distance from this centroid to all other foci's.
2. Sort the foci according to their distance from the centroid.
3. Make the points at the selected foci the scan window Z_i.
4. Include the points in the nearest foci to make the next scan window Z_{i+1}.
5. Continue including the points in the nearest foci until the prespecified proportion of points is reached.
6. Repeat 1 to 5 until all the foci are selected as centroids.

Once the scan region is defined, the computation of the test statistic and its corresponding significance is a straightforward task. The DCluster package in R (Gómez-Rubio, Ferrndiz-Ferragud, & Lopez-Qulez,

2005), provides a platform for implementing the above algorithm. The functions in the package may also be modified to accommodate test statistics other than the Kulldorff and Nagarwalla (1995) spatial scan statistic.

Example 1 To demonstrate our methods in application, we use the well known North Carolina SIDS data set to detect clusters of sudden infant death syndrome (SIDS) cases in the US state of North Carolina. The data is comprised of the number of live births, and the number of SIDS cases for each of the State's 100 county district in the two periods 1974-1978 and 1979-1984. Table 1 provides a summary of the data and the incidence of SIDS cases per 1000 live births.

Region	SIDS Cases	Live Births	Incidence Rate
A	139	36376	3.821
B	59	14388	4.101
North Carolina	1503	752354	1.998

Table 1: A summary of the data. Regions A and B are suspected clusters (see below).

This data was first presented by Symons, Grimson, and Yuan (1983). The spatial aspects of the data where considered progressively in the literature first by Cressie and Read (1985), and later expanded in Cressie and Chan (1989), and Cressie (1993). The data set along with all with its spatial attributes maybe found in the R spatial dependence package *spdep* (Bivand et al., 2011).

The coordinates of the county seats are used as the foci of the geographical grid. The measure at the foci, $\mu(Z)$, is the number of live births in the county, and the points $+$ of interest are the number of sids cases reported in the county. The scanning window is a cricle centered at the county seat and is extended to include neighboring counties in their entirety until the measure $\mu(Z)$ constitutes $R = 15\%$ of the total population.

Four tests for spatial cluster detection were applied: the Binomial, Fisher, and Normal tests presented in this paper and the Kulldorff (1997) spatial scan statistic. All four tests proved consistent results in that they all detected the same primary cluster, but with varying degrees of significance. The primary cluster consisted of the five counties of Bladen, Hoke, Columbus, Robeson and Scotland. We will identify this area as region A.

The search for clusters of secondary significance did not produce as consistent results. Depending on the test used, three regions, B, C, and D where identified. Both the Normal test and Kulldorff's spatial scan

	Primary Cluster			Secondary Cluster		
Test	Region	Test Statistic	p-value	Region	Test Statistic	p-value
Normal	A	7.98	$7.54 \times E^{-16}$	B	6.09	$5.46 \times E^{-10}$
Kulddorff	A	3046825	$9.90 \times E^{-06}$	B	273004.9	$5.00 \times E^{-04}$

Table 2: A summary of the clusters detected by the two test statistics.

statistic identify the region B as a secondary cluster. Region B consists of the counties Halifax, Hertford, and Northampton. The Binomial test detects a larger secondary cluster in the region C. The seven counties of Halifax, Hertford, Northampton, Bertie, Edgecombe, Gates, and Warren. It is worth noting that region B is a subregion of C and both regions share the same centriod. The Fisher test detects a cluster, region D, spanning 12 counties. Region D includes the counties of Beaufort, Bertie, Edgecombe, Greene, Halifax, Lenoir, Martin, Nash, Northampton, Pitt, Wayne, and Wilson. The centroid of D is different then that of B and C. The results are summarized in Table 2.

An obvious advantage of the tests proposed in this paper is that power of the test may be computed. The power function of the asymptotic Normal test statistic for the null hypothesis in $\mathcal{B}$, $\pi(\theta)$, is computed and shows that the test is quite powerful. For the SIDS example, $P_{Z|+}$ under the null hypothesis is $\nu(Z) = 0.04835$. The plot of the power function is shown in Figure 1.For this example, the probability of type I error is 0.049008 and this is achieved when the power function is evaluated at $\pi(\nu(Z) = 0.04835)$. As θ approaches 0, the variance of the asymptotic test statistic approaches 0. Hence, the test statistic z will become degenerate at 0 causing the power function $\pi(\theta) \to 0$ as $\theta \to 0$. From the plot we can see that as the value of θ surpasses $\nu(Z)$, the power of the test increases sharply.

Example 2 We use simulated data to compare the performance of the Kulldorff test and the asymptotic normal test given in this paper. For a fixed value of $\mu(G) = 5000$, $\mu(Z)$ is randomly generated from a uniform distribution with parameters $(0.5 * \mu(G), (R = 0.15) * \mu(G))$. We let $q = 0.2$, where q is the probability of a point $(+)$ in Z^c. The probability of a point, $+$, in Z, p, is allowed to diverge away from q to $p = 0.5$ by increments of 0.01. For each value of p the number of points in Z are simulated from a binomial distribution with parameters $(\mu(Z), p)$. Likewise, $n_+(G)$ is generated randomly from a binomial distribution with parameters $(\mu(G), q = 0.2)$. The p-values for the Kulldorff test (KT) and the asymptotic

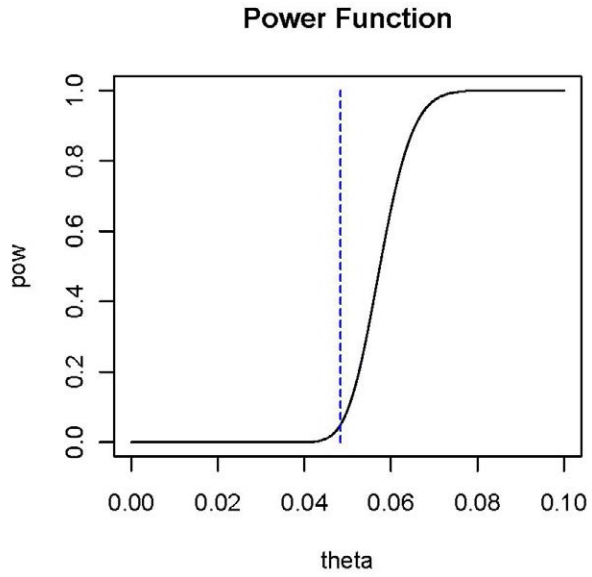

Figure 1: Power function for the test in Hypothesis $\mathcal{B}$. The dashed line marks the value of $\nu(Z)$.

normal test (ANT) are computed for each value of p. The results are depicted in Figure 2. We observe that the p-value of KT drops to significant levels only when p is noticeably greater then q. As depicted in the graph, this occurs when p is greater than 0.42. In contrast, the ANT p-value decreases sharply in response to the slightest deviation of p from q and returns consistent results as the gap widens. This is indicative that the ANT is a highly sensitive test. The KT requires a much larger discrepancy between p and q before this is reflected in the p-value.

6. Discussion

In this article we give an alternative view for the Kulldorff spatial scan testing hypotheses, and established test statistics in spatial scan statistics that are alternatives to the Kulldorff's test statistics. For large data set the test statistics with asymptotic normal distribution is as effective as the Kulldorff test statistics with the advantage that its limiting distribution is specified, which is valuable in analytical derivations in testing hypothesis, as in obtaining the power of the test.

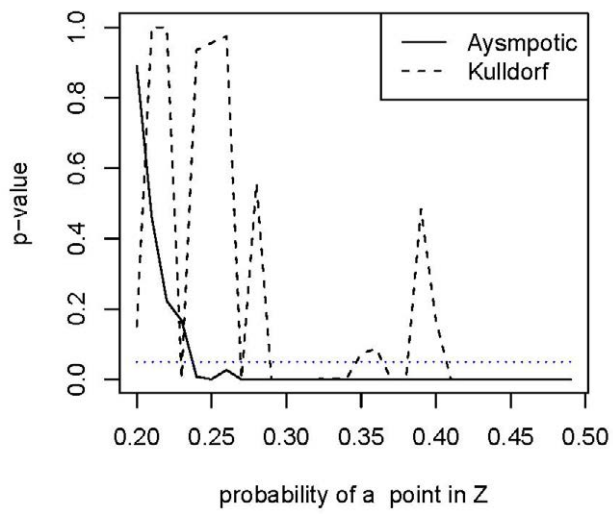

Figure 2: The p-values of KT and ANT for $0.2 \leq p \leq 0.5$ using simulated data. The horizontal dashed line marks the significance level, $\alpha = 0.05$.

References

Bivand, R., with contributions by Micah Altman, Anselin, L., Assuno, R., Berke, O., Bernat, A., ... Yu., D. (2011). spdep: Spatial dependence: weighting schemes, statistics and models [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=spdep` (R package version 0.5-31)

Cressie, N. (1993). *Statistics for spatial data*. J. Wiley. Retrieved from `http://books.google.com.kw/books?id=4SdRAAAAMAAJ`

Cressie, N., & Chan, N. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, *84*(406), 393-401.

Cressie, N., & Read, T. R. C. (1985). Do sudden infant deaths come in clusters? *Statistics and Decisions, Supplement Issue*, *3*(2), 333–349.

Glaz, J., & Naus, J. (1983). Multiple clusters on the line. *Communications in Statistics-Theory and Methods*, *12*(17), 1961-1986.

Gómez-Rubio, V., Ferrndiz-Ferragud, J., & Lopez-Qulez, A. (2005). Detecting clusters of disease with r.

Journal of Geographical Systems, *7*(2), 189-206.

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

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters - detection and inference. *Statistics in Medicine*, *14*(8), 799-810.

Loader, C. (1991). Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, *23*(4), 751-771.

Naus, J. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, *60*(310), 532-538.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International journal of geographical information systems*, *1*(4), 335-358.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`

Symons, M., Grimson, R., & Yuan, Y. (1983). Clustering of rare events. *Biometrics*, *39*(1), 193-205.

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, *4*(1), 11.

Turnbull, B., Iwano, E., Burnett, W., Howe, H., & Clark, L. (1990). Monitoring for clusters of disease - application to leukemia incidence in upstate New-York. *American Journal of Epidemiology*, *132*(1, S), S136-S143.

Zhang, T., & Lin, G. (2009). Spatial scan statistics in loglinear models. *Computational Statistics & Data Analysis*, *53*(8), 2851-2858.