COMMUN. STATIST.—THEORY METH., 26(6), 1481-1496 (1997)

A SPATIAL SCAN STATISTIC

Martin Kulldorff

Biometry Branch, DCPC, National Cancer Institute
EPN 344, 6130 Executive Blvd, Bethesda MD 20892, USA
and
Department of Statistics, Uppsala University, 751 20 Uppsala, Sweden

Keywords: point patterns; inhomogeneous Poisson process; Bernoulli process; clusters; clustering; confounders; sudden infant death syndrome.

ABSTRACT

The scan statistic is commonly used to test if a one dimensional point process is purely random, or if any clusters can be detected. Here it is simultaneously extended in three directions: (i) a spatial scan statistic for the detection of clusters in a multi-dimensional point process is proposed, (ii) the area of the scanning window is allowed to vary, and (iii) the baseline process may be any inhomogeneous Poisson process or Bernoulli process with intensity proportional to some known function. The main interest is in detecting clusters not explained by the baseline process.

These methods are illustrated on an epidemiological data set, but there are other potential areas of application as well.

1. INTRODUCTION

A scan statistic is used to detect clusters in a point process. It has been studied in the one-dimensional setting by Naus (1965a) and by many others. For a point process on an interval $[a, b]$, a window $[t, t + w]$ of fixed size $w < b - a$ is moved along the interval. Over all possible values of t, the maximum number

of points in the window is recorded and compared to its distribution under the null hypothesis of a purely random Poisson process.

The one-dimensional problem has been extended in various directions. When the points are grouped into one of several sub-intervals we have aggregated data. This has been studied by Wallenstein et al. (1989) among others, and is of interest when we have, for instance, monthly counts of some event. Weinstock (1981) has studied the problem where under the null-hypothesis the intensity of the underlying Poisson process has a known inhomogeneity. Various authors, such as Saperstein (1972) and Naus (1974), have studied a related Bernoulli model, with a sequence of binary outcomes. Loader (1991) allows for a non-fixed window size. Glaz and Naus (1983) have looked at a scan statistic searching for multiple clusters. For any of these extensions, and depending on the application, the scan statistic may or may not be conditioned on the total number of points observed.

In this paper a spatial scan statistic is proposed. An attempt is made to treat the problem in a setting as general as possible, except that the analysis is always conditioned on the total number of observed points. The window may take any predefined shape and the size of the window is allowed to vary as it scans the study region. The latter is very useful when we lack a prior knowledge about the size of the area covered by the cluster. The method also allows for an arbitrary, but known, underlying intensity that governs the distribution of points under the null hypothesis. This can take many different forms depending on the application. It is modeled as a measure μ on a geographical space G. When G is a line and μ is a uniform measure on $[a, b]$, we obtain the traditional one dimensional problem as a special case. With the Lebesgue measure on the plane we have a homogeneous spatial Poisson process. Other possible measures include the following:

1 The spatial clustering of trees is studied in forestry. A problem of potential interest is to see if there are clusters of trees that are of a specific kind or that have a certain characteristic, after having compensated for the uneven spatial distribution of all trees. That is, we want to know if the proportion of one kind of tree is particularly high in some location. The measure of an area would in this case be the total number of trees growing there.

2 In astronomy, there is an equivalent three-dimensional problem if we want to detect clusters of a particular kind of star after compensating for the irregular spatial distribution of all stars.

3 Epidemiologists are interested in geographical clusters of disease. Here it is necessary to compensate for the uneven density of the population as a whole. When data is aggregated into census districts the measure will be concentrated at the central coordinates of those districts.

4 To find uranium deposits, airplanes measure Geiger counts as they fly in parallel lines over large areas. A high number of counts in a specific area

indicates a possible deposit. The measure would be uniform along the flying lines and zero elsewhere.

5 A zoologist may study the spatial distribution of sea gull nests. In an archipelago, the nests will be located on islands. The appropriate measure would be uniform over land and zero elsewhere.

6 If we are interested in space time clusters of a disease, then the measure would still be concentrated in the geographical dimension as in example 3, but it would also be extended in a third dimension reflecting the population size as it changes over time in each of the census districts.

7 It is not always sufficient to adjust for an uneven population distribution, whether it be humans, trees, stars or something else. We may also need to take various confounders into account. For example, in epidemiology we could let the measure reflect the age standardized expected incidence rate.

In one dimension, the exact distribution of the test statistic is known only in special cases. Much of the literature has been concerned with finding good approximations. In higher dimensions the statistical theory becomes more complex. Naus (1965b) obtained distributional bounds for a two dimensional scan statistic on a square with uniform underlying measure, and with a rectangular window of fixed but arbitrary size. Loader (1991) treated the same problem but allowed for variable window size. Turnbull et al. (1990), using the underlying measure as described in example (3) above, used a circular window with constant measure. Since the exact distribution of the test statistic could not be determined, Monte Carlo simulation was used to perform the hypothesis test. All the models above are special cases of what is outlined in this paper. Two other special cases can be found in Kulldorff and Nagarwalla (1995) and Hjalmars et al. (1995) who apply the spatial scan statistic to data sets of leukemia in Upstate New York and Sweden respectively.

When the size of the scanning window is fixed, the test statistic is always taken as the maximum number of points in the window at any given time. With a variable window size that is no longer possible, and instead the likelihood ratio test statistic is used (Loader, 1991).

In section 2 the Poisson and the Bernoulli models are described. The likelihood ratio test statistic is then presented in Section 3, while Section 4 describes some of its theoretical properties. A discussion on computational issues and a practical example are given in Sections 5 and 6, respectively.

2. POISSON AND BERNOULLI MODELS

Let N denote a spatial point process where $N(A)$ is the random number of points in the set $A \subset G$. As the window moves over the study area it defines a

collection $\mathcal{Z}$ of zones $Z \subset G$. Interchangeably, Z will be used to denote both a subset of G and a set of parameters defining the zone.

For the Bernoulli model we consider only measures μ such that $\mu(A)$ is an integer for all subsets $A \subset G$. Each unit of measure corresponds to an 'entity' or 'individual' who could be in either one of two states, for example with or without some disease, or being of a certain species or not. Individuals in one of these states are defined as points, and the location of those individuals constitute the point process. In the model, there is exactly one zone $Z \subset G$ such that each individual within that zone has probability p of being a point, while the probability for individuals outside the zone is q. The probability for any one individual is independent of all the others. The null hypothesis is $H_0 : p = q$. The alternative hypothesis is $H_1 : p > q, Z \in \mathcal{Z}$. Under H_0, $N(A) \sim Bin(\mu(A), p)$ for all sets A. Under H_1 $N(A) \sim Bin(\mu(A), p)$ for all sets $A \subset Z$, and $N(A) \sim Bin(\mu(A), q)$ for all sets $A \subset Z^c$.

Under the Poisson model, points are generated by an inhomogeneous Poisson process. There is exactly one zone $Z \subset G$ such that $N(A) \sim Po(p\mu(A \cap Z) + q\mu(A \cap Z^c))$ $\forall A$. The null hypothesis is $H_0 : p = q$, while the alternative hypothesis states that $H_1 : p > q, Z \in \mathcal{Z}$. Under H_0, $N(A) \sim Po(p\mu(A))$ $\forall A$. Note that one of the parameters, Z, disappears under the null hypothesis. This is unusual but not unheard of, see for example Davies (1977).

The best choice of window, and thereby the corresponding collection $\mathcal{Z}$ of zones, depends on the application. Some possibilities are:

1 All circular subsets.

2 All circles centered at any of several foci on a fixed grid, with a possible upper limit on circle size. (Kulldorff and Nagarwalla, 1994)

3 Same as (2) but with a fixed circle size. (Turnbull et al., 1989)

4 All rectangles of a fixed size and shape. (Naus, 1965b)

5 When looking for space-time clusters we could use cylinders, scanning circular geographical areas over variable time intervals.

The underlying purpose of specifying a precise alternative is not to exclude other possibilities. On the contrary, the purpose is not only to detect clustering in the form of that specific alternative, but also to detect clusters in the form of other similar alternatives. For example, if a test has good power for an alternative with circular zones, then it will have fairly good power for squares and many ellipses as well, but not necessarily for a narrow zone stretching from one corner of a map to another.

How should we choose between the Bernoulli and the Poisson models? The choice does not matter much when the total number of points is small compared to $\mu(G)$. The Bernoulli and Poisson models then closely approximate each other. In other cases it will depend on the application. If we have binary

counts, such as two types of stars, then we should use the Bernoulli model. If we have counts relating to some continuous risk factor, as with a Geiger meter, we should use the Poisson model.

3. LIKELIHOOD RATIO TEST

It is now time to derive the likelihood ratio test. It is slightly different for the Bernoulli and the Poisson models, and we start with the former. Let n_Z denote the observed number of points in zone Z, and n_G the total number of observed points.

3.1 Bernoulli model

The likelihood function for the Bernoulli model is expressed as

$$L(Z,p,q) = p^{n_Z}(1-p)^{\mu(Z)-n_Z} q^{n_G-n_Z}(1-q)^{(\mu(G)-\mu(Z))-(n_G-n_Z)}$$

To detect the zone that is most likely to be a cluster, we find the zone $\hat{Z}$ that maximizes the likelihood function. In other words, $\hat{Z}$ is the maximum likelihood estimator of the parameter Z. We do this in two steps. First we maximize the likelihood function conditioned on Z.

$$L(Z) \stackrel{\text{def}}{=} \sup_{p>q} L(Z,p,q) = \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(1-\frac{n_Z}{\mu(Z)}\right)^{\mu(Z)-n_Z}$$

$$\times \left(\frac{n_G-n_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n_Z}\left(1-\frac{n_G-n_Z}{\mu(G)-\mu(Z)}\right)^{(\mu(G)-\mu(Z))-(n_G-n_Z)} \quad (1)$$

when $\frac{n_Z}{\mu(Z)} > \frac{(n_G-n_Z)}{(\mu(G)-\mu(Z))}$, and otherwise

$$L(Z) = \left(\frac{n_G}{\mu(G)}\right)^{n_G}\left(\frac{\mu(G)-n_G}{\mu(G)}\right)^{\mu(G)-n_G}$$

Next, we find the solution $\hat{Z} = \{Z : L(Z) \geq L(Z')\ \forall Z' \in \mathcal{Z}\}$. In Section 5, the numerical calculation of $\hat{Z}$ is discussed. The most likely cluster is of interest in itself, but we are also interested in making statistical inference. Let

$$L_0 \stackrel{\text{def}}{=} \sup_{p=q} L(Z,p,q) = \left(\frac{n_G}{\mu(G)}\right)^{n_G}\left(\frac{\mu(G)-n_G}{\mu(G)}\right)^{\mu(G)-n_G}$$

The likelihood ratio, λ, can be written as

$$\lambda = \frac{\sup_{Z\in\mathcal{Z},p>q} L(Z,p,q)}{\sup_{p=q} L(Z,p,q)} = \frac{L(\hat{Z})}{L_0} \quad (2)$$

Note that the denominator depends only on the total number of points n_G and not on the spatial distribution of the points. The ratio λ is used as the test statistic, and how to obtain its distribution through Monte Carlo replicas is described in Section 5.

So far we have discussed clusters in terms of an abnormally high number of cases in some area. It could be that we are instead interested in detecting areas with an unusually low number of points. This can be accomplished by simply changing the direction of the two inequality signs in equation 1. The same is true for the Poisson model.

3.2 Poisson model

The likelihood function for the Poisson model is a little more complex. The probability of n_G number of points in the study area is

$$\frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}[p\mu(Z)+q(\mu(G)-\mu(Z))]^{n_G}}{n_G!}.$$

The density function $f(x)$ of a specific point being observed at location x is

$$\begin{cases} \frac{p\mu(x)}{p\mu(Z)+q(\mu(G)-\mu(Z))} & \text{if } x \in Z \\ \frac{q\mu(x)}{p\mu(Z)+q(\mu(G)-\mu(Z))} & \text{if } x \notin Z \end{cases}$$

We can hence write the likelihood function as

$$L(Z,p,q) = \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}[p\mu(Z)+q(\mu(G)-\mu(Z))]^{n_G}}{n_G!}$$

$$\times \prod_{x_i \in Z} \frac{p\mu(x_i)}{p\mu(Z)+q(\mu(G)-\mu(Z))} \prod_{x_i \notin Z} \frac{q\mu(x_i)}{p\mu(Z)+q(\mu(G)-\mu(Z))}$$

$$= \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}}{n_G!} p^{n_Z} q^{(n_G-n_Z)} \prod_{x_i} \mu(x_i) \qquad (3)$$

As before, the likelihood ratio is defined as in equation 2. We have

$$L_0 = \sup_p \frac{e^{-p\mu(G)}p^{n_G}}{n_G!} \prod_{x_i} \mu(x_i) = \frac{e^{-n_G}}{n_G!} \left(\frac{n_G}{\mu(G)}\right)^{n_G} \prod_{x_i} \mu(x_i)$$

For the numerator we first take the supremum over all p and q for a fixed Z. Equation 3 takes its maximum when $p = n_Z/\mu(Z)$ and $q = (n_G - n_Z)/(\mu(G) - \mu(Z))$, so

$$L(Z) = \begin{cases} \frac{e^{-n_G}}{n_G!} \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(\frac{n_G-n_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n_Z} \prod_{x_i} \mu(x_i) & \\ & \text{if } \frac{n_Z}{\mu(Z)} > \frac{(n_G-n_Z)}{(\mu(G)-\mu(Z))} \\ \frac{e^{-n_G}}{n_G!} \left(\frac{n_G}{\mu(G)}\right)^{n_G} \prod_{x_i} \mu(x_i) & \\ & \text{otherwise.} \end{cases}$$

The test statistic λ of the likelihood ratio test can now be written as

$$\lambda = \frac{\sup_{Z \in \mathcal{Z}} L(Z)}{\frac{e^{-n_G}}{n_G!}\left(\frac{n_G}{\mu(G)}\right)^{n_G}\prod_{x_i}\mu(x_i)}$$

$$= \sup_{Z \in \mathcal{Z}} \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_G - n_Z}{\mu(G)-\mu(Z)}\right)^{n_G - n_Z}}{\left(\frac{n_G}{\mu(G)}\right)^{n_G}} I\left(\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}\right) \qquad (4)$$

if there is at least one zone Z such that $\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}$, and $\lambda = 1$ otherwise. I() is the indicator function.

4. PROPERTIES OF THE TEST STATISTIC

4.1 Detection versus inference

Most statistical methods for cluster analysis of a spatial point process are either descriptive in the sense that they can detect the location of clusters but without any inference involved, or they do inference but without the ability to detect the location of clusters. An important characteristic of the spatial scan test is that it does both, so that when the null hypothesis is rejected we can locate the specific area of the map that causes the rejection. To be precise, let $\mathbf{x} = \{x_i, i = 1, .., n_G\}$ denote the set of coordinates of the n_G points in a data set where $\hat{Z}$ is the most likely cluster, and let $\mathbf{x}' = \{x'_i, i = 1, .., n_G\}$ be an alternative configuration with exactly the same number of points. The following theorem holds for the Bernoulli and Poisson models.

Theorem 1 *If the null hypothesis is rejected under* $\mathbf{x}$ *then it is also rejected under* $\mathbf{x}'$, *if* $x'_i = x_i$ *for all* $x_i \in \hat{Z}$.

In words, the theorem states that as long as the points within the zone constituting the most likely cluster are located where they are, we would still reject the null hypothesis no matter how the rest of the points were shuffled around. For example, if the null hypothesis is rejected due to a disease cluster in Seattle, it does not matter how we move around the cases on the U.S. east coast, the null-hypothesis will still be rejected. This might sound like a self evident property, but it does not hold for most other tests for spatial clustering such as Knox (1964), Whittemore et al. (1987), Cuzick and Edwards (1990), or Diggle and Chetwynd (1991). Those tests are hence not suitable if we want to know the location of clusters. Rather, they are geared towards answering

the question of whether the phenomenon of clustering occurs over the study region as a whole, such as if a disease is infectious or not, a question for which the spatial scan statistic is not suitable.

Proof: Let $\lambda(\mathbf{x})$ and $\lambda(\mathbf{x}')$ denote the values of the test statistic for the two different data sets. Since the two data sets have the same number of points, the distribution of λ under the null hypothesis will be the same, and it is hence enough to show that $\lambda(\mathbf{x}') \geq \lambda(\mathbf{x})$. In the Bernoulli case we have

$$\lambda(\mathbf{x}) = \frac{L(\hat{Z}|\mathbf{x})}{L_0} \leq \frac{L(\hat{Z}|\mathbf{x}')}{L_0} \leq \frac{\sup_Z L(Z|\mathbf{x}')}{L_0} = \lambda(\mathbf{x}').$$

The first inequality holds since $\mathbf{x}'$ has at least as many points within zone Z as $\mathbf{x}$. For the Poisson model it is trivially true if $\lambda(\mathbf{x}) = 1$. When $\lambda(\mathbf{x}) > 1$ we have from equation 4 that

$$\begin{aligned}
\lambda(\mathbf{x}) &= \sup_Z \frac{1}{K}\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_G - n_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n_Z} \\
&= \frac{1}{K}\left(\frac{n_{\hat{Z}}}{\mu(\hat{Z})}\right)^{n_{\hat{Z}}}\left(\frac{n_G - n_{\hat{Z}}}{\mu(G)-\mu(\hat{Z})}\right)^{n_G-n_{\hat{Z}}} \\
&\leq \frac{1}{K}\left(\frac{n'_{\hat{Z}}}{\mu(\hat{Z})}\right)^{n'_{\hat{Z}}}\left(\frac{n_G - n'_{\hat{Z}}}{\mu(G)-\mu(\hat{Z})}\right)^{n_G-n'_{\hat{Z}}} \\
\leq \sup_Z \frac{1}{K}\left(\frac{n'_Z}{\mu(Z)}\right)^{n'_Z}&\left(\frac{n_G - n'_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n'_Z} = \lambda(\mathbf{x}').
\end{aligned}$$

where $K = (n_G/\mu(G))^{n_G}$. The first inequality holds since for any constants α, β and N, $(\alpha n)^n(\beta(N-n)^{N-n}$ is an increasing function of n when $\alpha n > \beta(N-n)$.

4.2 Power

The power of the one dimensional scan statistic has been studied by Wallenstein et al. (1993,1994) and Sahu et al. (1993) among others. For the spatial scan statistic we cannot expect to find a uniformly most powerful test, except in the special case when there is only one zone in the alternative hypothesis. Instead, we show that it fulfills a criterion making it what we call an *individually most powerful* test.

To define an individually most powerful test we divide the composite alternative hypothesis into distinct subsets. The parameter space is partitioned into a countable number of subsets $\{A_j\}$ such that $A_j \cap A_{j'} = \emptyset$ for all $j \neq j'$, and such that $\cup A_j$ constitutes the whole of the alternative hypothesis. Likewise, and using the same index, the critical region C, where the null hypothesis is rejected, is partitioned into disjoint subsets $\{C_j\}$ where $\cup C_j = C$. Let $C' = \cup C'_j$ denote an alternative critical region with corresponding disjoint subsets.

Definition 1 *For a particular significance level α, a test is* individually most powerful *with respect to a partition $\{A_j\}$ of the parameter space, and a partition $\{C_j\}$ of the critical region, if for each A_k there are no sets C' and $\{C'_j\}$ such that*

1. $C_j = C'_j$ *for all* $j \neq k$.
2. $P(\omega \in C'|H_0) = \alpha$
3. $P(\omega \in C'_k|(Z,p,q)) > P(\omega \in C_k|(Z,p,q))$ *for any* $(Z,p,q) \in A_k$.

This means that if we fix the critical region except for its subset C_k as indicated by statement 1, then the test is uniformly most powerful compared to all remaining choices of the critical region and with respect to all parameters $(Z,p,q) \in A_k$. This property is very important in any multiple testing type of a situation, where there is a composite alternative hypothesis and where we wish to know which part of it causes the rejection. As mentioned before, the scan statistic has the ability to identify the zone responsible for rejecting the null hypothesis, and if we fail to detect a real cluster, it is of little comfort if the null hypothesis is rejected based on an untrue cluster in another part of the study area. In fact, that is usually less desirable than just failing to reject the null hypothesis. The problem resembles other multiple comparison situations, where instead of testing multiple cluster locations simultaneously, we might test several new agricultural crop varieties to see if any of them are better than the one presently in use, or we might simultaneously test several potential risk factors for cancer.

If we are only concerned about rejection versus no rejection, without an interest in the location of clusters, then the property of being an individually most powerful test is of little value. For such a problem the likelihood ratio based spatial scan statistic would be a suboptimal choice.

Now, let $A_Z = \{(Z,p,q) : p > q\}$ and $A_0 = \{(Z,p,q) : p = q\}$. Let C_Z denote the intersection of the critical region C and the subset of the sample space in which Z is the most likely cluster.

Theorem 2 *The test based on λ forms an individually most powerful test with respect to the partitions $\{A_Z\}$ and $\{C_Z\}$. This holds for the Bernoulli as well as the Poisson model.*

Proof: We show that if statements (1) and (2) in the definition are true, then (3) cannot hold. For an arbitrary Z, let $D_- = \{\omega : \omega \in C_Z, \omega \notin C'_Z\}$ and $D_+ = \{\omega : \omega \in C'_Z, \omega \notin C_Z\}$. Let

$$M = \sup_{\omega \in D_+} \frac{L(Z,p,q|\omega)}{L(H_0|\omega)}.$$

For the Poisson model, it follows from equation 4 that

$$M = \sup_{\omega \in D_+} \frac{\left(\frac{n_Z(\omega)}{\mu(Z)}\right)^{n_Z(\omega)} \left(\frac{n_G - n_Z(\omega)}{\mu(G) - \mu(Z)}\right)^{n_G - n_Z(\omega)}}{(n_G/\mu(G))^{n_G}} \leq$$

$$\inf_{\omega \in D_-} \frac{\left(\frac{n_Z(\omega)}{\mu(Z)}\right)^{n_Z(\omega)} \left(\frac{n_G - n_Z(\omega)}{\mu(G) - \mu(Z)}\right)^{n_G - n_Z(\omega)}}{(n_G/\mu(G))^{n_G}} = \inf_{\omega \in D_-} \frac{L(Z,p,q|\omega)}{L(H_0|\omega)}$$

A similar argument holds for the Bernoulli case, based on equation 1. Now, for any $(Z,p,q) \in A_Z$,

$$\begin{aligned}
&P(\omega \in C'_Z|(Z,p,q)) - P(\omega \in C_Z|(Z,p,q)) \\
&= P(\omega \in D_+|(Z,p,q)) - P(\omega \in D_-|(Z,p,q)) \\
&= \int_{\omega \in D_+} p(\omega|Z,p,q)d\omega - \int_{\omega \in D_-} p(\omega|Z,p,q)d\omega \\
&= \int_{\omega \in D_+} L(Z,p,q|\omega)d\omega - \int_{\omega \in D_-} L(Z,p,q|\omega)d\omega \\
&= \int_{\omega \in D_+} \frac{L(Z,p,q|\omega)}{L(H_0|\omega)} L(H_0|\omega)d\omega - \int_{\omega \in D_-} \frac{L(Z,p,q|\omega)}{L(H_0|\omega)} L(H_0|\omega)d\omega \\
&\leq \int_{\omega \in D_+} ML(H_0|\omega)d\omega - \int_{\omega \in D_-} ML(H_0|\omega)d\omega \\
&= M\left(\int_{\omega \in D_+} P(\omega|H_0)d\omega - \int_{\omega \in D_-} P(\omega|H_0)\right)d\omega \\
&= M\left(P(\omega \in D_+|H_0) - P(\omega \in D_-|H_0)\right) \\
&= M\left(P(\omega \in C'_Z|H_0) - P(\omega \in C_Z|H_0)\right) \\
&= M\left(P(\omega \in C'|H_0) - P(\omega \in C|H_0)\right) = 0
\end{aligned}$$

The second to last equality holds since $C_j = C'_j$ for all $j \neq Z$ according to statement 1 in the definition.

5. COMPUTATIONS AND MONTE CARLO SAMPLING

In order to find the value of the test statistic, we need a way to calculate the likelihood ratio as it is maximized over the collection of zones in the alternative hypothesis. This might seem like a daunting task since the number of zones could easily be infinite. Two properties allows us to reduce it to a finite problem. The number of observed points is always finite and for a fixed number of points the likelihood decreases as the measure of the moving window increases.

Consider the scanning window of example (2) in Section 2. If we let the circle size increase for a fixed foci, we only need to recalculate the likelihood whenever a new point enters the circle. Since there is a finite number of points, the number of times we need to compute the likelihood for each foci is finite, and since the number of foci is also finite, the total number of calculations is finite. Assuming a Bernoulli model or a homogeneous Poisson model, similar arguments hold for the other four examples given in Section 2.

Once the value of the test statistic has been calculated, it is easy to do the inference. We cannot expect to find the distribution of the test statistic in closed analytical form. Instead we rely on Monte Carlo simulation. Originally proposed by Dwass (1957), this technique was first used in the context of a scan statistic by Turnbull et al. (1990). Because we know the underlying measure μ, we can obtain replications of the data set generated under the null hypothesis when we condition on the total number of points n_G. With 9999 such replications, the test is significant at the 5 percent level if the value of the test statistic for the real data set is among the 500 highest values of the test statistic coming from the replications.

In addition to the most likely cluster we might also want to look at secondary clusters with high likelihood values. Some of these will be related to the most likely cluster in the sense that they contain about the same set of points with their respective zones overlapping each other. Such secondary clusters are usually of little interest, although they serve to remind us of the fact that the obtained location and size of detected clusters are only estimates.

More interesting types of secondary clusters are those located in another part of the study region. We define these to be clusters that do not overlap with a more likely one. It is often of interest to report these clusters along with the most likely one.

For inference, we may take a secondary cluster and compare and rank its likelihood value with the maximum likelihood ratio from the Monte Carlo replications. Any secondary cluster that ranks below the significance level would in itself have caused the rejection of the null hypothesis even if there had been no other more likely cluster in the data set. This gives us an inferential procedure for secondary clusters as well, but since we are comparing a secondary cluster from the data set with the most likely clusters from the replications such a test is somewhat conservative.

6. EXAMPLE

We illustrate the models using data on sudden infant death syndrome (SIDS) in North Carolina. The data were compiled by M. Symmons, D. Atkinson,

TABLE I

The spatial scan statistic applied to sudden infant death syndrome in North Carolina, adjusted for the uneven geographical distribution of births. Zones refer to Figure 1 and incidence is the number of deaths per 1000 live births.

	Zone Z	# SIDs n_Z	# Births $\mu(Z)$	Incidence	p-value
Bernoulli	A	139	36376	3.8	0.0001
model	B	59	14388	4.1	0.0005
Poisson	A	139	36376	3.8	0.0001
model	B	59	14388	4.1	0.0003

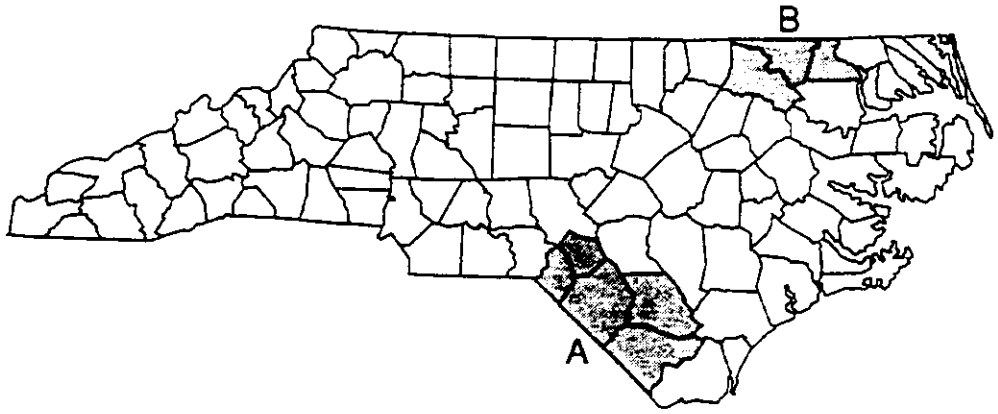

FIG 1: Two significant clusters of sudden infant death syndrome in North Carolina, adjusted for the uneven geographical distribution of live births.

and the State Center for Health Statistics of the North Carolina Department of Human Resources. They have previously been analysed by Cressie and Chan (1989) among others.

For each of the 100 counties in North Carolina, the data comprise the total number of live births as well as the number of sudden infant deaths (SIDs) for the years 1974-1984. The number of live births in the counties ranges from 567 to 52345. The location of county seats were used as the geographical coordinates. The total number of SIDs are 1503 out of 753354 live births. This gives a state wide incidence rate of 2.0 per 1000. The total

number of births in each county, as well as the statewide number of SIDs, are also stratified into whites and non-whites. The complete data are presented by Cressie and Chan (1989).

The measure at the coordinate point of each county is taken as the number of live births in that county. The measure is zero elsewhere. This is as in example (3) of Section 1. As zones for the window we use all circles that are centered at one of the county coordinate points and that include at most half of the total population. This follows example (2) of Section 2.

Note that the zones are circular only with respect to the aggregated data. As we draw the circles around one county seat, other counties will either be completely part of a zone or else not at all, depending on whether its county seat falls within the circle or not. Hence, we get a compact but irregular shaped zone following the county boundaries. This can be seen in Figures 1 and 2.

6.1 Bernoulli Model

The Bernoulli model is the most natural one to use for this data set. We have birth counts, and each birth can correspond to at most one sudden infant death. Table 1 summarizes the results of the analysis.

The most likely cluster, A, consists of the counties of Bladen, Columbus, Hoke, Robeson, and Scotland, in the southern part of the state. The rank is 1/10000, i.e. a p value of 0.0001.

There is one other significant cluster, B, composed of Halifax, Hertford, and Northampton counties in the northeast. With a rank of 5/10000 it has a p value of 0.0005. This latter test is conservative, because we are comparing a secondary cluster in the data set with the most likely clusters from the replicas.

6.2 Poisson Model

Since we are dealing with a rare disease, the Poisson model should give a close approximation to the Bernoulli model. That the results are indeed similar for this data set can be seen in Table 1.

The Poisson approximation is especially useful when we have covariates that we wish to include in the analysis. For SIDS, one possible covariate is race (Cressie and Chan, 1989), which may be related to SIDS through unobserved variables such as quality of housing or access to health care. The racial distribution differs widely among the counties in North Carolina, and could possibly explain the previously detected clusters. We may want to see if there are still geographic clusters after adjusting for race. This could lead us to other spatially related risk factors that are otherwise hidden.

The overall incidence of SIDS is 1.512 for white children, and 2.970 for non-white children (Cressie and Chan, 1989). The underlying measure at each county coordinate x can now be defined as

TABLE II

The spatial scan statistic applied to sudden infant death syndrome in North Carolina, adjusted for race and the uneven geographical distribution of live births. Zones refer to Figure 2.

	Zone Z	# SIDs n_Z	E[# SIDs] $\mu(Z)$	# Births	p-value
Poisson	A	139	94.5	36376	0.0036
model	C	191	140.8	86780	0.0060

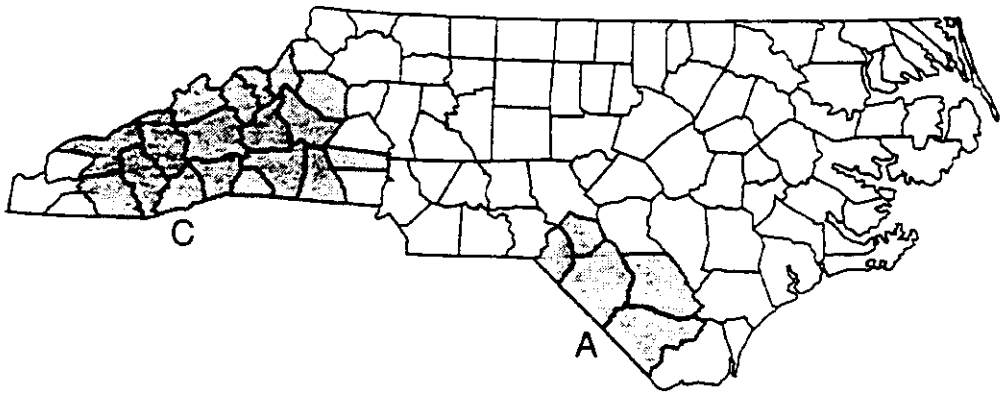

FIG 2: Two significant clusters of sudden infant death syndrome in North Carolina, adjusted for race and the uneven geographical distribution of live births.

$$\mu(x) = \text{white births} \times 1.512 + \text{nonwhite births} \times 2.970$$

which is proportional to the expected number of SIDs under the null hypothesis. Note that we do not need to know the number of SIDs in each county subdivided by race. The result of the likelihood ratio test is given in Table 2.

Comparing this analysis to the one where race was not incorporated, we observe three things.

1 With a rank of 36/10000 ($p = 0.0036$), the southern cluster A remains significant, and cannot be explained solely by the high proportion of non-white births in that area.

2 Cluster B in the northeast is no longer significant, with a rank of 3336/10000 ($p = 0.3336$).

3 A previously 'hidden' cluster C emerges in the west, with a rank of 60/10000 ($p = 0.006$). It consists of the following counties: Avery, Buncombe, Burke, Caldwell, Cleveland, Haywood, Henderson, Jackson, Lincoln, Macon, Madison, Mitchell, Polk, Rutherford, Swain, Transylvania, and Yancey.

ACKNOWLEDGEMENTS

Valuable discussions with Laurence Freedman and Lisa McShane are gratefully acknowledged. This research was partly funded by the Swedish Research Council in the Humanities and Social Sciences.

BIBLIOGRAPHY

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

Cuzick J and Edwards R, (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society* Ser. B, 52, 73-104.

Davies RB, (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247-254.

Diggle PJ and Chetwynd AG, (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* 47, 1155-1163.

Dwass M, (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28, 181-187.

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

Hjalmars U, Kulldorff M, Gustafsson G and Nagarwalla N, (1996). Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine* 15, 707-715.

Knox G, (1964). The detection of space-time interactions. *Applied Statistics* 13, 25-29.

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

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

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

Naus JI, (1965b). Clustering of random points in two dimensions. *Biometrika* 52, 263-267.

Naus J, (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association* 69, 810-815.

Sahu SK, Bendel RB and Sison CP, (1993). Effect of relative risk and cluster configuration on the power of the one-dimensional scan statistic. *Statistics in Medicine* 12, 1853-1865.

Saperstein B, (1972). The generalized birthday problem. *Journal of the American Statistical Association* 67, 425-428.

Turnbull BW, Iwano EJ, Burnett WS, Howe HL and Clark LC, (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 132, S136-S143.

Wallenstein S, Weinberg CR and Gould M, (1989). Testing for a pulse in seasonal event data. *Biometrics* 45, 817-830.

Wallenstein S, Naus J and Glaz J, (1993). Power of the scan statistic for detection of clustering. *Statistics in Medicine* 12, 1829-1843.

Wallenstein S, Naus J and Glaz J, (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence. *Biometrika* 81.

Weinstock MA, (1981). A generalized scan statistic test for the detection of clusters. *International Journal of Epidemiology* 10, 289-293.

Whittemore AS, Friend N, Brown BW and Holly EA, (1987). A test to detect clusters of disease. *Biometrika* 74, 631-635, and 75, 396.

Received September, 1996; Revised December, 1996.