14

Spatial Scan Statistics: Models, Calculations, and Applications

Martin Kulldorff

National Cancer Institute, Bethesda, MD

Abstract: A common problem in spatial statistics is whether a set of points are randomly distributed or if they show signs of clusters or clustering. When the locations of clusters are of interest, it is natural to use a spatial scan statistic.

Different spatial scan statistics have been proposed. These are discussed and presented in a general framework that incorporates two-dimensional scan statistics on the plane or on a sphere, as well as three-dimensional scan statistics in space or in space–time. Computational issues are then looked at, presenting efficient algorithms that can be used for different scan statistics in connection with Monte Carlo-based hypothesis testing. It is shown that the computational requirements are reasonable even for very large data sets. Which scan statistic to use will depend on the application at hand, which is discussed in terms of past as well as possible future practical applications in areas such as epidemiology, medical imaging, astronomy, archaeology, urban and regional planning, and reconnaissance.

Keywords and phrases: Spatial statistics, geography, spatial clusters, space–time clusters, maximum likelihood, likelihood ratio test

14.1 Introduction

The scan statistic is a statistical method with many potential applications, designed to detect a local excess of events and to test if such an excess can reasonably have occurred by chance. The scan statistic was first studied in detail by Naus (1965a,b), who looked at the problem in both one and two dimensions.

J. Glaz et al. (eds.), *Scan Statistics and Applications*

In two or more dimensions, which is the topic of this chapter, the events may be cases of leukemia, with an interest to see if there are geographical clusters of the disease; they may be antipersonnel mines, with an interest to detect large mine fields for removal; they could be Geiger counts, with an interest to detect large uranium deposits; they could be stars or galaxies; they could be breast calcifications showing up in a mammography, possibly indicating a breast tumor; or they could be a particular type of archaeological pottery. Later on we will discuss each of these and several other applications and the type of scan statistic that is suitable in each situation.

Three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. We present a general framework in which most multidimensional scan statistics fit. Depending on the application, different models will be chosen, and depending on the model, the test statistic may be evaluated either through explicit mathematical derivations and approximations or through Monte Carlo sampling. In the latter case, random data sets are generated under the null hypothesis, and the scan statistic is calculated in each case, comparing the values from the real and random data sets to obtain a hypothesis test.

While computer intensive, the Monte Carlo approach need not be overly so. In this chapter, we present a set of efficient algorithms which can be used to calculate the spatial scan statistic for a set of different models with a circular window. One of these with a continuously variable radius, required 163 minutes of computing time on a 100 MHz Pentium PC, when applied to 65,040 cases of melanoma in the 3,053 counties of the continental United States.

Section 14.2 is essentially a review of the existing literature, while Section 14.3 presents mostly new material. Section 14.4 describes how the spatial scan statistic can be utilized in practice in an attempt to inspire its use in current as well as new areas of application.

14.2 Models

14.2.1 A general model

As mentioned above, the three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window.

Kulldorff (1997) defined a general model for the multidimensional scan statistic. Let A be the area in which events may occur, a subset of Euclidean space where different dimensions may represent either physical space or time.

For example, A could be particular geographical area during a ten-year period, where events are recorded both geographically and temporally.

On A define a measure μ, representing a known underlying intensity that generates events under the null hypothesis. For a homogeneous Poisson process on a rectangle A, we have $\mu(x) = \lambda$ for all $x \in A$ and some constant λ. The measure could also be discrete, so that it is only positive on a finite number of *population points*, where $\mu(B)$ is the combined measure of the population points located in area $B \subset A$. We require that $\mu(B) > 0$ for all areas B.

Let X denote a spatial point process where $X(B)$ is the random number of events in the set $B \subset A$. Two different probability models are considered, based on Bernoulli counts and the Poisson process, respectively.

For the Bernoulli model, we consider only discrete measures μ such that $\mu(B)$ is an integer for all subsets $B \subset A$. 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 events, and the location of those individuals constitute the point process. Under the null hypothesis, the number of events in any given area is binomially distributed, so that $X(B) \sim Bin(\mu(B), p)$ for some value p and for all sets $B \subset A$.

For the Poisson model, events are generated by a homogeneous or nonhomogeneous Poisson process. Under the null hypothesis, $X(B) \sim \text{Poisson}(p\mu(B))$, for some value p and for all sets $B \subset A$. The measure μ may either be defined continuously so that events may occur anywhere, or discretely so that events may occur only at prespecified locations, or as a combination of the two. The discrete case is useful when we are dealing with individual counts or with aggregated data.

The window of a scan statistic is often thought of as an interval, area, or volume of fixed size and shape, which then moves across the study area. As it moves, it defines a collection $\mathcal{W}$ of zones $W \subset A$. To be more general, we allow for windows of variable size and shape, by defining the window as a collection $\mathcal{W}$ of zones $W \subset A$ of any size and shape. What defines it as a scan statistic is that the different zones overlap each other and jointly cover the whole area A.

Conditioning on the observed total number of events, $X(A)$, the definition of the scan statistic is the maximum likelihood ratio over all possible zones

$$S_{\mathcal{W}} = \frac{\max_{W \in \mathcal{W}} L(W)}{L_0} = \max_{W \in \mathcal{W}} \frac{L(W)}{L_0} , \tag{14.1}$$

where $L(W)$ is the likelihood function for zone W, expressing how likely the observed data are given a differential rate of events within and outside the zone, and where L_0 is the likelihood function under the null hypothesis.

Let $X(A \setminus W) = X(A) - X(W)$ and $\mu(A \setminus W) = \mu(A) - \mu(W)$. For the Bernoulli model,

$$\frac{L(W)}{L_0} =$$

$$\frac{\left(\frac{X(W)}{\mu(W)}\right)^{X(W)} \left(1-\frac{X(W)}{\mu(W)}\right)^{\mu(W)-X(W)} \left(\frac{X(A\setminus W)}{\mu(A\setminus W)}\right)^{X(A\setminus W)} \left(1-\frac{X(A\setminus W)}{\mu(A\setminus W)}\right)^{\mu(A\setminus W)-X(A\setminus W)}}{\left(\frac{X(A)}{\mu(A)}\right)^{X(A)} \left(1-\frac{X(A)}{\mu(A)}\right)^{\mu(A)-X(A)}} \tag{14.2}$$

if $X(W)/\mu(W) > X(A \setminus W)/\mu(A \setminus W)$, and $L(W) = 1$ otherwise. For the Poisson model,

$$\frac{L(W)}{L_0} = \frac{\left(\frac{X(W)}{\mu(W)}\right)^{X(W)} \left(\frac{X(A\setminus W)}{\mu(A\setminus W)}\right)^{X(A\setminus W)}}{\left(\frac{X(A)}{\mu(A)}\right)^{X(A)}} \tag{14.3}$$

if $X(W)/\mu(W) > X(A\setminus W)/\mu(A\setminus W)$, and $L(W) = 1$ otherwise. The expression $X(W)/\mu(W) > X(A \setminus W)/\mu(A \setminus W)$ simply states that there are more than the expected number of events within the window as compared to outside the window. If we were scanning for areas with a low number of events, then ">" would change to "<." For details and derivations as a likelihood ratio test, see Kulldorff (1997), who has also proved some optimal properties for these test statistics.

When the window size is fixed in terms of the expected number of events, that is, if $\mu(W) = \mu(W')$ for all $W, W' \in \mathcal{W}$, then the scan statistic is

$$S'_{\mathcal{W}} = \max_{W\in\mathcal{W}} X(W),$$

the maximum number of events in the window over all possible locations. Note that $S'_{\mathcal{W}} \neq S_{\mathcal{W}}$, but for any two realizations of the point process, say ω_1 and ω_2, $S'_{\mathcal{W}}(\omega_1) > S'_{\mathcal{W}}(\omega_2)$ if and only if $S_{\mathcal{W}}(\omega_1) > S_{\mathcal{W}}(\omega_2)$. This means that, when the window size is fixed, then a hypothesis test based on $S'_{\mathcal{W}}$ is identical to one based on $S_{\mathcal{W}}$.

For a Poisson model with continuous measure, a lower bound on the window size is needed. If not, then a window containing a sequence of increasingly smaller zones all containing the same event will in the limit give an infinite valued test statistic. It is also natural to put an upper bound on the window size. A window W that contains almost all of A makes little sense, and should be interpreted as a lack of events outside of W rather than as an excess inside.

14.2.2 Special cases

Both one and multidimensional scan statistics are special cases of the above model. Many features of it originated in connection with one-dimensional scan statistics; see, for example, Saperstein (1972), Naus (1974), Weinstock (1981), Wallenstein, Weinberg, and Gould (1989b), and Glaz and Naus (1991). Here, we review the multi-dimensional literature.

In terms of the area A being scanned, Naus (1965b), Loader (1991), Alm (1997, 1998) and Anderson and Titterington (1997) all considered a rectangle. Alm (1998) also looked at a three-dimensional rectangular volume. Chen and Glaz (1996) looked at a regular grid of discrete points within a rectangular area. Turnbull *et al.* (1990) used an irregular grid, where points may be anywhere within an arbitrarily shaped area.

Under the null hypothesis, Naus (1965b), Loader (1991), and Alm (1997, 1998) looked at a homogeneous Poisson process, Turnbull *et al.* (1990) considered a nonhomogeneous Poisson process, while Anderson and Titterington (1997) considered both types. Chen and Glaz (1996) considered a Bernoulli model.

As for the scanning window, Naus (1965b), Loader (1991), Chen and Glaz (1996), Alm (1997, 1998) and Anderson and Titterington (1997) all considered rectangles. In addition, Alm (1997, 1998) also looked at circles, triangles, and other convex shapes. Turnbull *et al.* (1990) considered a circular window centered at any of the grid points making up the data. The window is, in all cases, of fixed shape as well as of fixed size in terms of the expected number of events, with the exception of Loader (1991), who also considered a variable size window.

In terms of applications, the general model has been applied in a number of different settings, the first of which was presented at the SPRUCE conference in 1992 and later published by Kulldorff and Nagarwalla (1995). For all of these, the data are located on an irregular grid within an arbitrarily shaped area. Kulldorff and Nagarwalla (1995) and Section 6.1 of Kulldorff (1997) used the Bernoulli model, while Section 6.2 of Kulldorff (1997), Hjalmars *et al.* (1996), Kulldorff *et al.* (1997, 1998), and Walsh and Fenster (1997) used a nonhomogeneous Poisson process. In terms of the scanning window, all used a variable size circle centered on the grid points, except for Kulldorff *et al.* (1998), who used a three-dimensional cylinder where the size of both the base and the height is variable independently of each other.

The choice of scan statistic will depend on the particular application at hand, a topic we will turn to in Section 14.4.

14.2.3 Related methods

As part of a "geographical analysis machine," Openshaw *et al.* (1987) used a number of overlapping circular zones of different radii. The purpose is the same as with a spatial scan statistic, to detect clusters of events, but a separate test is performed for each of the many zones. This leads to multiple testing, and even under the null hypothesis we would expect a large number of "significant" clusters, but as a descriptive geographical analysis tool the method is useful. Turnbull *et al.* (1990) solved the problem of the multiple testing for circles with fixed expected number of cases, while Kulldorff and Nagarwalla (1995)

and Kulldorff (1997) solved it for variable size circles.

Priebe (1998) proposed a spatial scan statistic for stochastic scan partitions. In a two-step procedure, one set of data is first used to create a set of non-overlapping zones, called scan partitions, while another set of data containing the events is used to see if any of these partitions have a statistically significant excess of events. Because the zones are nonoverlapping, the calculations for the second part are more simple than for a standard scan statistic. It is necessary to have the additional data set though, used in the first step, and under the null hypothesis the two data sets need to be independent of each other for the test to be valid.

In other related problems, Eggleton and Kermack (1944), Besag and Newell (1991), Månsson (1996), and many others have studied the number of clusters of some prespecified magnitude. Lawson (1997) applied a Bayesian framework to investigate the number of clusters and their locations. Adler (1984), Worsley *et al.* (1992), and some others have investigated the supremum of a Gaussian random field.

Wallenstein, Gould, and Kleinman (1989a) used a scan statistic in the time dimension to improve on a previously proposed space–time clustering test, but the test itself is not a scan statistic. Rather than taking a maximum over the geographical zones, the degree of clustering in each zone is summed over all zones, making it a global clustering test. Such tests are useful for quite different purposes, when the locations of clusters are not of interest.

14.3 Calculations

14.3.1 Probabilistic approximations

The mathematics for obtaining the distribution of the scan statistic is quite complex, and exact derivations have proved elusive for all but the simplest scenarios. There are some very interesting and impressive probabilistic approximations though. Starting with Naus (1965b), later results have been obtained by Loader (1991), Chen and Glaz (1996), and Alm (1997,98). Månsson (1996) has derived some limit results. Details of these developments can be found in Chapter 5 of this volume by Sven-Erick Alm, and in Chapter 10 by Marianne Månsson.

14.3.2 Monte Carlo-based hypothesis testing

When probabilistic approximations are not available, Monte Carlo-based hypothesis testing is. In principle, this can be applied to any special case of the general model presented in Section 14.2. Generating random cases is typically

not a problem, but calculating the value of the test statistic can be a complex undertaking, depending on the model chosen. For the descriptive cluster detection method described earlier, Openshaw *et al.* (1987) used a Cray supercomputer even though their approach is conceptually simpler than a scan statistic. By using efficient statistical algorithms, the calculation times can be substantially reduced.

Monte Carlo-based hypothesis testing was proposed by Dwass (1957), who pointed out that the probability of falsely rejecting the null hypothesis is exactly according to the significance level, in spite of the simulation involved. Mantel (1967) proposed its use in terms of spatial point processes, while Turnbull *et al.* (1990) was the first to use it in the context of a multidimensional scan statistic. Monte Carlo hypothesis testing for a scan statistic is a four-step procedure:

1. Calculate the value of the test statistic for the real data.

2. Create a large number of random data sets generated under the null hypothesis.

3. Calculate the value of the test statistic for each of the random replications.

4. Sort the values of the test statistic, from the real and random data sets, and note the rank of the one calculated from the real data set. If it is ranked in the highest α percent, then reject the null hypothesis at α percent significance level.

The key in terms of minimizing computing time is Step 3, as it can be complex in nature, and most of all, because it must be repeated once for each random replication of the data set. Anderson and Titterington (1997) presented the following algorithm for a circular window of fixed diameter d on a homogeneous Poisson process:

Algorithm 14.3.1 (Anderson-Titterington: Circular window. Fixed size. Homogeneous Poisson process.)

1. *Identify the locations (x, y) of two events no more than distance d apart.*

2. *Construct the two circles of diameter d for which x and y lie on the circumference.*

3. *Identify the number of events that lie on or inside each of the two circles and let n be the larger of those two numbers.*

4. *Repeat Steps 1 to 3 for all relevant pairs of locations and report the largest of the resulting n-values as being the scan statistic.*

The complexity of one visit to Step 3 is of the order $O(N)$, where $N = X(A)$, the total number of events. Steps 1–3 must be repeated $O(N^2)$ times for each of R Monte Carlo replications, so the total complexity is $O(RN^3)$. When N is large, a more efficient algorithm is:

Algorithm 14.3.2 (Circular window. Fixed size. Homogeneous Poisson process.)

1. *Identify the location x of an event and construct a large circle with radius d centered at x. Pick an arbitrary location on the large circle, x_0, and denote the angle from x to x_0 as $0°$.*

2. *Create a smaller circle of radius $d/2$ within the larger one. Imagine the smaller circle moving clockwise completely within the larger circle in such a way that x is always on its circumference. Denote by x_a, $0° < a < 360°$, the single point that is on the circumference of both circles, where a is the angle from x to x_a.*

3. *For each event on or inside the larger circle, note the two angles of the line from x to x_a when the event enters and departs the smaller moving circle. Sort the angles in increasing order, keeping track of whether the angle corresponds to an entrance or a departure.*

4. *For the smaller circle which has both x and x_0 on its circumference, count the number of events inside it. Then go through the array of sorted angles from $0°$ to $360°$, adding one to the count for each entrance, subtracting one for each departure. Denote the maximum count by n.*

5. *Repeat Steps 1 to 4 for all events, and report the largest of the resulting n-values as the scan statistic.*

6. *Repeat Steps 1 to 5 for each Monte Carlo replication.*

Each visit to Steps 2 and 4 is $O(N)$ while the sorting in Step 3 is $O(NlogN)$. There are N iterations of Steps 1 to 5 for each of R replications, and hence the total complexity is $O(RN)[O(N) + O(NlogN) + O(N)] = O(RN^2logN)$.

In most practical applications, the cluster size is unknown a priori. For a homogeneous Poisson process, the simplest algorithm to program would be to pick all triplets of events, in turn, and for each triplet construct the circle for which all three events lie on the circumference, then counting the number of events within that circle. Based on the number of events and the circle size, it is then possible to calculate the likelihood according to (14.3), and the largest likelihood over all possible triplets is the scan statistic. Such an algorithm is $O(RN^4)$. A more efficient algorithm, with complexity $O(RN^3logN)$, is as follows.

Algorithm 14.3.3 (Circular window. Variable size. Homogeneous Poisson process.)

1. *Identify the locations* (x, y) *of two events, and construct the straight line* L *between the two where each point on the line is equal distance from* x *and* y. *Denote one end of the line as the left end.*

2. *For each remaining event* z, *construct the circle such that all of* (x, y, z) *lie on the circumference. Note where on* L *lies the circle centroid corresponding to* z, *and whether event* z *enters or departs the circle as the centroid moves toward the left.*

3. *Sort the circle centroids on* L *from right to left, keeping track of whether that centroid corresponds to an entrance or a departure.*

4. *Calculate the number of events in the circle with its centroid farthest to the right, as well as the circle size. Then move down the sorted array of circles centroids adding or subtracting events as they enter or depart the circle. For each circular area* W, *register the number of events* n *as well as the circle measure* $\mu(W) = \int_W \mu \, dy = \mu\pi r^2$, *where* r *is the radius.*

5. *Repeat Steps 1 to 4 for all pairs of events, and report the largest likelihood based on all* $(n, \mu(W))$*-pairs as the scan statistic, where the likelihood is calculated according to (14.3).*

6. *Repeat Steps 1 to 5 for each Monte Carlo replication.*

So far, we have presented algorithms for homogeneous Poisson processes. A simple case of a nonhomogeneous Poisson process is a gradual linear shift in intensity so that $\mu(x) = a + bx$ for some a and b. Algorithm 14.3.3 can be easily modified to account for this by calculating the measure of the circular area W centered at x as $\mu(W) = \int_W \mu(y)dy = \mu(x)\pi r^2 dx$, where r is the circle radius.

Another form of nonhomogeneity is the discrete case in which the measure is concentrated on a finite set of population points. The following algorithm is similar to Algorithm 14.3.3 but based on the location of the population points containing positive measure, rather than on the location of events. We can no longer calculate the measure simply from the circle size, and hence, we need to keep track of the amount of measure in the window simultaneously with the number of events.

Algorithm 14.3.4 (Circular window. Variable size. Discrete nonhomogeneous Bernoulli or Poisson process.)

1. *Identify the locations* (x, y) *of two population points, and construct the straight line* L *between the two where each point on the line is equal distance from* x *and* y. *Denote one end of the line as the left end.*

2. *For each remaining population point z, construct the circle such that all of (x, y, z) lie on the circumference. Note where on L lies the circle centroid and whether the population point enters or departs the circle as the centroid moves toward the left.*

3. *Sort the circle centroids located on L from right to left, keeping track of whether that centroid corresponds to an entrance or departure.*

4. *Calculate the number of events n in the circle with its centroid farthest to the right on the line, as well as the measure $\mu(W)$ for that circle. Then move down the sorted array of circles centroids adding and subtracting events and measure as population points enter and depart the circle. For each circular area W, register the number of events n as well as the population measure $\mu(W)$.*

5. *Repeat Steps 1 to 4 for all pairs of population points, and report the largest likelihood based on all $(n, \mu(W))$-pairs as the scan statistic, where the likelihood is calculated according to (14.2) in the case of a Bernoulli model, and according to (14.3) for the Poisson model.*

6. *Repeat Steps 1 to 5 for each Monte Carlo replication.*

The complexity of this algorithm is $O(RM^3 logM))$, where M is the number of population points.

For most applications, it is not crucial to include all possible circles in the set of zones constituting the window, and an alternative is to use only a subset of closely overlapping circles. This reduces the computing time. In the following two algorithms, the window contains only those circles that are centered at any of a number of prespecified irregular grid points. The radius of the circles still vary continuously.

Algorithm 14.3.5 (Circular window. Variable size. Circle centroids on grid. Homogeneous Poisson process.)

1. *Pick a grid point. Calculate the distance to the different events and sort in increasing order.*

2. *Create a circle centered at the grid point and continuously increase the radius. For each event entering the circle, note the number of events n and the measure $\mu(W) = \mu\pi r^2$ inside the circle.*

3. *Repeat Steps 1 and 2 for each grid point. Report the largest likelihood based on all $(n, \mu(W))$-pairs as the scan statistic, where the likelihood is calculated according to (14.3).*

4. *Repeat Steps 1 to 3 for each Monte Carlo replication.*

The complexity of this algorithm is $O(RGNlogN)$, where G is the number of grid points. For a discrete nonhomogeneous process, we have the following:

Algorithm 14.3.6 (Circular window. Variable size. Circle centroids on grid. Discrete nonhomogeneous Bernoulli or Poisson process.)

1. *Pick a grid point. Calculate the distance to the different population points and sort those in increasing order. Memorize the sorted population points in an array.*
2. *Repeat Step 1 for each grid point.*
3. *Pick a grid point.*
4. *Create a circle centered at the grid point and continuously increase the radius. For each population point entering the circle, update the number of events n and the measure $\mu(W)$ inside the circular area W.*
5. *Repeat Steps 3 and 4 for each grid point. Report the largest likelihood based on all $(n, \mu(W))$-pairs as the scan statistic, where the likelihood is calculated according to (14.2) or (14.3).*
6. *Repeat Steps 3 to 5 for each Monte Carlo replication.*

The complexity of Steps 1 and 2 is $O(GMlogM)$, as this does not have to be repeated for each Monte Carlo replication. The complexity of Steps 3 to 6 is $O(RGM)$.

Algorithms 14.3.5 and 14.3.6 also work for three-dimensional spherical windows by simply defining the population points in three-dimensional space. The complexity remains the same but for a complete coverage, the number of grid points G may have to be larger.

In space–time applications, one option is simply to define time as a third dimension and use a spherical window on that three-dimensional space. One problem with this is that the result will depend on the relative units of spatial and temporal distances. Another problem is that a sphere would represent a cluster starting with zero spatial size, then growing steadily over time until a maximum spatial size is reached, after which it gradually shrinks back to zero size again. It is more natural to scan for clusters using the intersection of a spatial circle and a temporal interval, leading to a cylindrical window. Algorithm 14.3.6 can be adjusted for this purpose, if for each geographical circle, we also scan the time-dimension using a variable size temporal interval. It also means that the geographical and temporal size can vary independently of each other. The complexity of Steps 3 to 6 then becomes $O(RGMN^2)$ if exact times are known, and $O(RGMI^2)$ if times are aggregated into I time intervals.

Algorithms 14.3.1 to 14.3.6 extend to circular windows on the surface of a sphere, by simply defining the events and population points in three dimensions

on the spherical surface, and by adjusting the calculations of circle sizes and distances accordingly. This is very useful for geographical applications, avoiding the need for two-dimensional map projections.

Scanning for low rates can also be handled by any of the mentioned algorithms. For Algorithms 14.3.4 and 14.3.6, it is just a question of changing the sign of the inequality when calculating the likelihood $L(W)$. For the other algorithms, it is also necessary to subtract the number of events on the border of the circle from the circle total.

A circular window has the advantage of being invariant under a rotation of the space. There are applications though where other shapes are of interest. Anderson and Titterington (1997) gave an $O(RN^2)$ algorithm for a square window of fixed size with sides parallel to the axes of the coordinate system. A scan statistic with a fixed shape variable size ellipsoidic window can be calculated using any of Algorithms 14.3.3 to 14.3.6, by rescaling one of the axes in the underlying coordinate system. We leave it for future research to present algorithms for other models.

14.3.3 Software

For certain multidimensional scan statistics, Kulldorff and Williams (1997) have developed *SaTScan*. This software is available free of charge from the authors, or from the World Wide Web at `http://dcp.nci.nih.gov/BB/SaTScan.html`.

SaTScan uses Algorithm 14.3.6, and is based on a nonhomogeneous Poisson process defined on an irregular grid; it can be used to analyze the following types of multidimensional scan statistics: (i) a scan statistic on the plane with a circular window of variable size with centroids on an arbitrarily defined regular or irregular grid, (ii) same on the surface of a sphere such as the earth, (iii) a three-dimensional scan statistic with variable size spheric windows centered on an arbitrary irregular grid, (iv) a space–time scan statistic with a variable size cylinder, where the base of the circle corresponds to a geographical area, and the height to a time interval, and where the sizes of the circle and interval are variable independently of each other.

The software will, in all cases, adjust for any number of covariates specified by the user, and it is possible to scan for areas with a large number of events as well as for areas with a low number of events. Certain one-dimensional scan statistics can also be analyzed by putting all data on a single line. A future version will also include the Bernoulli model.

Using SaTScan, the calculations for the New Mexico example below took 8 seconds on a 100 MHz Pentium PC. With 1175 events in only 32 census areas, it is a rather small data set though. For the same type of analysis but with 65,040 cases of melanoma in the United States, aggregated to 3053 counties, SaTScan used 163 minutes of computer time when the maximum window size was set to 50% of the total, and 80 minutes when it was set to 10%. For 1592 cases of

leukemia in 2507 Swedish parishes, it used 62 and 15 minutes, respectively. The cylinder based space–time scan statistic used 21 hours for the Swedish data set with years as the temporal unit and 10% as the maximum geographic window size. The number of Monte Carlo replications were in all cases 999.

This shows that the computational requirements for the spatial scan statistic is quite reasonable in practical applications with very large data sets.

14.4 Applications

14.4.1 Epidemiology

There is a long history of geographical surveillance of disease by publishing disease atlases. If there are areas with exceptionally high rates, they may give us clues to the etiology of the disease, it may indicate areas where health care needs improvement, or it may indicate areas to be targeted for preventive measures. In those atlases that are not purely descriptive, analysis is often done by dividing the study region into nonoverlapping districts, making a separate test of hypothesis for each district to see if it has an excess incidence or mortality [Choynowski (1959)]. With a spatial or space–time scan statistic, we can do the surveillance adjusting for the multiplicity of possible cluster locations, without being limited by the boundaries of prespecified districts, and without defining the size of potential cluster a priori.

Events may be cases diagnosed of some disease or deaths due to that disease. The measure is by nature nonhomogeneous, reflecting the geographical distribution of the population at risk. In most situations, we want to adjust for covariates that are known risk factors such as age or sex. We might have individual locations for cases and all non-cases, or of cases and a random set of controls, but more often the data are aggregated at some small geographical level such as census tracts, parishes or postal code areas. In either case, we can use Algorithm 14.3.4 or 14.3.6.

If the population at risk is all births and the events are occurrences of sudden infant death syndrome [Kulldorff (1997)] or birth defects, then we should use the Bernoulli model. If on the other hand, we are looking at fatal cardiac arrest in a population, we choose the Poisson model since such individuals are no longer part of the population numbers after the event occurs. Most applications fall somewhere in between the two, but whenever the number of events is small compared to the population at risk, the two models approximate each other so that either could be chosen.

In terms of practical epidemiological applications, the spatial scan statistic has been used to study leukemia in Upstate New York by Turnbull *et al.* (1990) using a fixed size window, and by Kulldorff and Nagarwalla (1997) using a vari-

able size window. Hjalmars *et al.* (1996) have looked at childhood leukemia incidence in Sweden, Kulldorff (1997) studied sudden infant deaths in North Carolina, Kulldorff *et al.* (1997) have looked at breast cancer mortality in the northeastern United States, while Walsh and Fenster (1997) have studied mortality from systemic sclerosis in the southeastern United States. All these use a variable size circular window. Using a fixed size square window, Anderson and Titterington (1997) looked at laryngeal cancer in South Lancashire, England. The space–time scan statistic, using a variable size cylindrical window, has been applied to brain cancer incidence in New Mexico by Kulldorff *et al.* (1998).

14.4.2 Example: Brain cancer in New Mexico

To give an example, we look at the geographical distribution of brain cancer incidence in New Mexico. In 1989, a local resident detected an excess of brain cancer in Los Alamos during the previous year. This cluster alarm was evaluated statistically by Kulldorff *et al.* (1998) using a space–time scan statistic, without finding a significant space–time cluster in Los Alamos. Here, we will use a purely spatial scan statistic in more of a surveillance setting.

Broken down by age and sex, brain cancer and population data are available from 1973 to 1992 at the aggregated level of 32 counties. A circular variable size window was used. The circle centroids are limited to the county centroids, while the radius varies continuously from zero and up until it includes 50% of the total population at risk. Using a Poisson model, the analysis is adjusted for age and sex. One analysis was done scanning for areas with high rates (clusters) and another scanning for areas with low rates.

When scanning for areas with high rates, a cluster was found in and around Albuquerque, containing Bernadillo, Cibola-Valencia, Los Alamos, Sandoval, San Miguel, Santa Fe, Socorro, and Torrance counties (Figure 1.1), almost half the total state population. With 642 cases when 583.2 were expected, this area had a rate 10 percent higher than the New Mexico average, and it is significant with $p = 0.030$. As the New Mexico mortality rate was 16 percent lower than the United States average during 1986–90 [Miller *et al.* (1993)], this cluster may indicate that the Albuquerque area is more similar to the rest of the United States in terms of brain cancer than other parts of New Mexico.

When scanning for areas with low rates, the likelihood took on its maximum value for Lea and Eddy counties combined (Figure 14.1). With 72 cases when 97.4 were expected, these counties had an incidence rate 26 percent lower than the state average, with $p = 0.221$, a nonsignificant result.

When interested in areas with either high or low rates, then we can either do two one-sided tests as we have done above, or we can do a single two-sided test, which is recommended. The clusters found will be the same, but not the p-value. For the two-sided test, $p = 0.067$.

Note from Figure 14.1 that the detected clusters are not perfect circles even

though we used a circular window. This is because the data are aggregated to the county level, so that all of a county is considered to be within the window when the centroid is, and vice versa. The only way to obtain perfect circles is to have non-aggregated data.

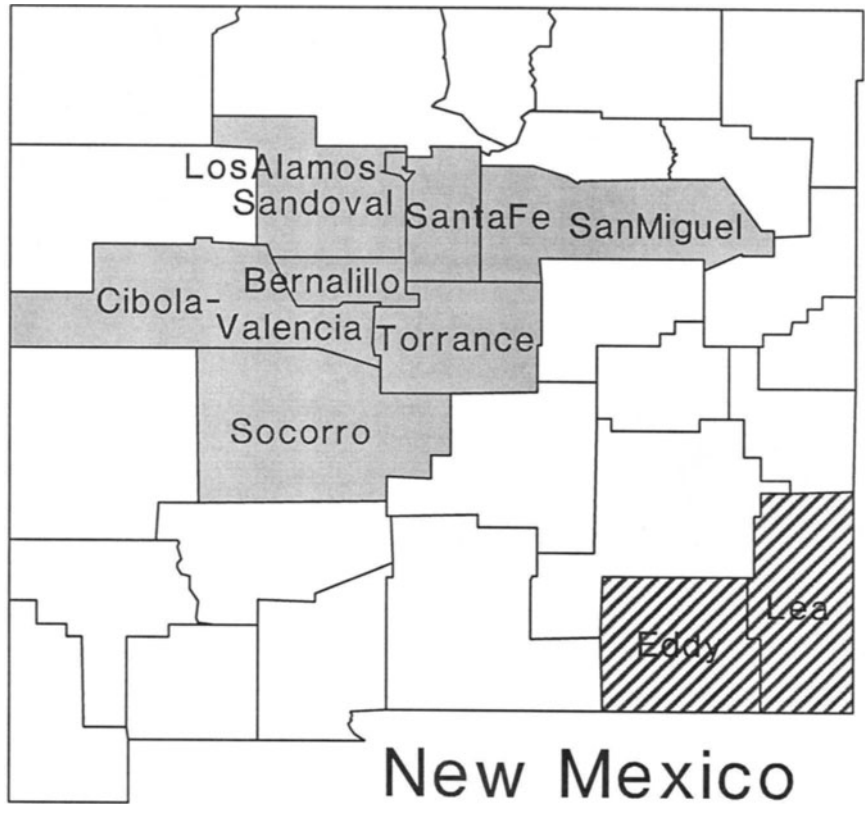

Figure 14.1: Brain cancer incidence in New Mexico 1973–1991: The most likely cluster around Albuquerque in Bernadillo county ($p = 0.030$) and the most likely area with exceptionally low rate in Lea and Eddy counties ($p = 0.221$)

14.4.3 Medical imaging

In medical imaging, the aim may be to detect tumors using mammography, or areas of activation in a brain scan related to certain physical or mental activities. There are applications in both two and three dimensions. Priebe (1998) applied his scan statistic based on random scan partitions on mammography images, looking for clusters of breast calcifications, and using the texture of the breast to define the scan partitions. Worsley *et al.* (1992) and others have looked at

the supremum of a Gaussian random field to determine centers of activity in the brain. The multidimensional scan statistic is a complementary approach to these problems, where each specific application will determine the best method to use.

14.4.4 Astronomy

The three-dimensional scan statistic can be used for two different types of astronomy problems. We could be interested to see if stars, galaxies, or some other type of heavenly object are randomly distributed, or whether there are significant local clusters. This leads to a homogeneous Poisson model and Algorithm 14.3.5. It can also be of interest to know whether a particular type of star or galaxy is randomly distributed after adjusting for the locations of all stars/galaxies. Then we should use the nonhomogeneous Bernoulli model and Algorithm 14.3.6.

14.4.5 Archaeology and history

Alt and Vach (1991) studied the location of graves containing individual, with a certain genetically determined odontological feature, comparing them to the locations of all graves within a prehistoric burial site. The purpose was to see if biologically related persons, who are more likely to share the same odontological feature, were buried close to each other. Using a test for global clustering, their main purpose was to test for spatial correlation without any interest in cluster locations. If we are interested in the latter, we would instead use a spatial scan statistic based on a discrete Bernoulli model with calculations based on Algorithms 14.3.4 or 14.3.6.

Other potential archaeological and historical applications include the geographical distribution of a certain type of pottery as compared to the distribution of all discovered pottery, to locate areas where that type is significantly abundant, the geographical location of cities or castles in relation to the population distribution, or the geographical distribution of villages with a certain name ending as compared to the distribution of all villages.

14.4.6 Urban and regional planning

Post offices, elementary schools, voting locations and many other establishments need to be fairly spread out so they can be conveniently reached by most people. By applying the spatial scan statistic to look for areas with an exceptionally low number of them, adjusting for the underlying population distribution, we may find underserved populations where additional localizations are warranted. Businesses could also use such an approach to help determine appropriate locations for restaurants, grocery stores, health clubs, hairdressers, etc.

14.4.7 Reconnaissance

Antipersonnel mines injure thousands of people each year long after the war for which they were intended has ended. It is of great importance to detect mines so they can be deactivated and removed. It is possible to scan a large area for possible mines from the air, but of the point locations obtained, only some will reflect true mines while others will be false detections. By using a scan statistic, areas most likely to contain mines can be detected.

For such an application, we have a homogeneous Poisson process under the null hypothesis. As the size of possible minefields are hard to know a priori, we should use a variable size window, leading to Algorithms 14.3.3 or 14.3.5.

If there are boundary features in the landscape in such a way that it is unlikely that a minefield would cut across such borders, then it is advantageous to use those to create scan partitions as suggested by Priebe (1998). That will increase the power of the test.

Another type of reconnaissance for which a spatial scan statistic can be useful is when searching for mineral, oil, or uranium deposits.

14.4.8 Power

Wallenstein, Naus, and Glaz (1993, 1994a,b) have provided simple approximations for the power of the one-dimensional scan statistic against a rectangular pulse alternative, and Sahu, Bendel, and Sison (1993) have shown that it has good power against other pulse alternatives such as triangles. This may indicate that multidimensional scan statistics also have good power against pulse alternatives, but that has never been thoroughly investigated. For one special case, it has been confirmed by Kulldorff and Nagarwalla (1995) who compared their model using a variable window size with the fixed window size model used by Turnbull *et al.* (1990). The variable size model had good power irrespective of the true cluster size. The fixed size model had higher power if the specified size was within about 20 percent of the true cluster size. Neither model had a problem detecting a square shaped cluster even though both used a circular window.

References

1. Adler, R. J. (1984). The supremum of a particular Gaussian field, *Annals of Probability*, **12**, 436–444.

2. Alm, S. E. (1997). On the distribution of the scan statistic of a two dimensional Poisson process, *Advances in Applied Probability*, **29**, 1–16.

3. Alm, S. E. (1998). On the distribution of scan statistics for Poisson processes in two and three dimensions, *Extremes* (to appear).

4. Alt, K. W. and Vach, W. (1991). The reconstruction of 'genetic kinship' in prehistoric burial complexes—problems and statistics, In *Classification, Data Analysis, and Knowledge Organization* (Eds., H. H. Bock and P. Ihm), Berlin: Springer-Verlag.

5. Anderson, N. H. and Titterington, D. M. (1997). Some methods for investigating spatial clustering with epidemiological applications, *Journal of the Royal Statistical Society, Series A*, **160**, 87–105.

6. Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A*, **154**, 143–155.

7. Chen, J. and Glaz, J. (1996). Two dimensional discrete scan statistics, *Statistics & Probability Letters*, **31**, 59–68.

8. Choynowski, M. (1959). Maps based on probabilities, *Journal of the American Statistical Association*, **54**, 385–388.

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

10. Eggleton, P. and Kermack, W. O. (1944). A problem in the random distribution of particles, *Proceedings of the Royal Society, Edinburgh Section*, **62**, 103–115.

11. Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

12. 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.

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

14. Kulldorff, M., Athas, W. F., Feuer, E. J., Miller, B. A. and Key, C. R. (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, *American Journal of Public Health* (submitted).

15. Kulldorff, M., Feuer, E. J., Miller, B. A. and Freedman, L. S. (1997). Breast cancer clusters in Northeast United States: A geographic analysis, *American Journal of Epidemiology*, **146**, 161–170.

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

17. Kulldorff, M. and Williams, G. (1997). *SaTScan v 1.0, Software for the Space and Space-Time Scan Statistics*, Bethesda, MD: National Cancer Institute.

18. Lawson, A. (1997). Cluster modeling of disease incidence via MCMC methods, *Journal of Statistical Planning and Inference* (submitted).

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

20. Månsson, M. (1996). On Clustering of Random Points in the Plain and in Space, *Ph.D. Thesis*, Department of Mathematics, Chalmers University of Technology and Gothenburg University, Gothenburg.

21. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Research*, **27**, 209–220.

22. Miller, B. A., Gloeckler Ries, L. Y., Hankey, B. F., Kosary, C. L., Harras, A., Devesa, S. S. and Edwards, B. K. (1993). *SEER Cancer Statistics Review 1973–1990*, Bethesda, MD: National Cancer Institute.

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

24. Naus, J. (1965b). Clustering of random points in two dimensions, *Biometrika*, **52**, 263–267.

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

26. Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987). A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems*, **1**, 335–358.

27. Priebe, C. (1998). A spatial scan statistic for stochastic scan partitions, *Journal of the American Statistical Association* (to appear).

28. Sahu, S. K., Bendel, R. B. and Sison, C. P. (1993). Effect of relative risk and cluster configuration on the power of the one-dimensional scan statistic, *Statistics in Medicine*, **12**, 1853–1865.

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

30. Turnbull, B., Iwano, E. J., Burnett, W. S., Howe, H. L. and Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in Upstate New York, *American Journal of Epidemiology*, **132**, S136–S143.

31. Wallenstein, S., Gould, M. S. and Kleinman, M. (1989a). Use of the scan statistic to detect time-space clustering, *American Journal of Epidemiology*, **130**, 1057–1064.

32. Wallenstein, S., Weinberg, C. R. and Gould, M. (1989b). Testing for a pulse in seasonal event data, *Biometrics*, **45**, 817–830.

33. Wallenstein, S., Naus, J. and Glaz, J. (1993). Power of the scan statistic for detection of clustering, *Statistics in Medicine*, **12**, 1819–1843.

34. Wallenstein, S., Naus, J. and Glaz, J. (1994a). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence, *Biometrika*, **81**, 595–601.

35. Wallenstein, S., Naus, J. and Glaz, J. (1994b). Power of the scan statistics, *ASA Proceedings of the Section of Epidemiology*, **81**, 70–75.

36. Walsh, S. J. and Fenster, J. R. (1997). Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, 1981-90, *Journal of Rheumatology* (to appear).

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

38. Worsley, K. J., Evans, A. C., Marrett, S. and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain, *Journal of Cerebral Blood Flow and Metabolism*, **12**, 900–918.