MULTISCALE ADVANCED RASTER MAP ANALYSIS SYSTEM
Geographical Surveillance for Hotspot Detection, Delineation, and Prioritization:
Spatial Scan Statistics for Irregularly Shaped Clusters and Early Warning System

G. P. Patil
Center for Statistical Ecology and Environmental Statistics
Department of Statistics
The Pennsylvania State University
University Park, PA 16802
http://www.stat.psu.edu/~gpp

Research and Outreach Team

G. P. Patil, Department of Statistics, The Pennsylvania State University
Martin Kulldorff, Department of Community Medicine, University of Connecticut
Ann C. Klassen, Health Policy & Management, Johns Hopkins School of Public Health
R. T. Heffernan, Communicable Disease Program, New York City Department of Health
C. Taillie, Department of Statistics, The Pennsylvania State University
R. M. Assunção, Department of Statistics, Federal University of Minas Gerais, Brazil
L. Duczmal, Department of Statistics, Federal University of Minas Gerais, Brazil

May 30, 2002
Research Plan

A. Specific Aims

In geographical disease surveillance, an important problem is the identification and/or evaluation of areas with exceptionally high or low prevalence, incidence or mortality. Such information can provide important clues about geographically related variables that predict occurrence and the course of the disease. The predictors of interest are not restricted to unknown etiological risk factors, such as occupational exposures, dietary habits or the natural environment. Equally important are various disease control factors such as mammography screening, access to health care and current treatment regimens.

The spatial scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) has become a popular method for detection and evaluation of disease clusters, and is now used by many health departments and academic epidemiologists both nationally and internationally (Appendix 1). In terms of its mathematical definition, the spatial scan statistic is defined in very general terms, in order to accommodate clusters of different shapes and forms. In practice, it is typically the circular spatial scan statistic that is used, due to circle’s natural compactness, available software (SaTScan) and computational efficiency. The circular scan statistic works well even when true cluster is not a circle, as long as it is reasonably compact. If the true cluster is very irregular though, such as along a long winding river or along a narrow urban corridor, a circular scan statistic has low power.

Specific Aim 1: To develop new versions of spatial scan statistic designed for detection of clusters of irregular shapes. We will consider three general approaches: an elliptic spatial scan statistic, a tree-structured spatial scan statistic and a simulated annealing based spatial scan statistic. The former looks for clusters of elliptic shapes while the latter two consider more general sets of connected census areas as the set of potential clusters.

One of the most important uses of the scan statistic for disease surveillance is in the early detection of disease outbreaks. For this purpose, the circular spatial scan statistic has been replaced by a cylinder based space-time scan statistic. This is for example used daily by the New York City Health Department for hospital emergency admissions based syndromic surveillance to detect any emerging health problems as early as possible.

Specific Aim 2: Each of the three methods described in specific aim 1 will be generalized to space-time versions for the early detection of disease outbreaks.

For certain problems, there is an underlying network structure on which we want to perform the cluster detection and evaluation. For example, the New York City Health Department is monitoring the New York subway system and water distribution networks for bioterrorism attacks. In such a scenario, a circular scan statistic is not useful as two individuals close to each other in Euclidian distance may be very far from each other along the network.

Specific Aim 3: The latter two methods will be employed for the detection and evaluation of clusters on a predefined network.

As with any new methods development, it is important to compare methods and test them and modify them based on real practical applications.
Specific Aim 4: The three methods will be compared with each other as well as with the circular spatial scan statistic, looking at statistical power, geometric accuracy in cluster estimation, and computing time. More than a million simulated benchmark data sets are already available for this purpose (Kulldorff et al 2002). These will be augmented with data sets displaying a broader range of true cluster shapes.

Specific Aim 5: All methods will be tested and applied on real data sets. These include prostate cancer incidence and staging data from the Maryland Cancer Registry, breast cancer mortality data from Northeastern United States, and syndromic surveillance data from the New York City Subway System.

B. Background and Significance

B1. Geography of Disease and Disease Surveillance
Disease maps can provide important clues concerning geographical variability in the etiology, prevention, screening and treatment of disease. Disease atlases are often produced for this reason, most commonly for cancer (e.g., Devesa et al 1999; Le et al 1996) but also for many other health events such as heart disease (Casper et al 1999; Barnett et al 2001), teenage births (Taylor et al 1996) and different types of mortality (Pickle et al 1996). Individual disease maps are also commonly produced for specific diseases in medical journal articles, health department reports and disease registry publications.

Most disease registries are updated at least on a yearly basis. If a geographically localized health hazard suddenly occurs, we would like to have a surveillance system in place that can pick up a new geographical disease cluster as quickly as possible, irrespective of its location and size. At the same time, we want to minimize the number of false alarms. Other types of health data such as hospital emergency room admissions, 911 calls and pharmacy calls are available on a daily basis, and are now starting to be used by health departments for syndromic surveillance in order to detect disease outbreaks as early as possible. A system that provides a warning only a few days earlier can have enormous benefits to public health when an outbreak occurs, as disease control and prevention measures can be activated at an earlier stage. This is especially important for unusual but potentially very serious infectious diseases.

B2. The Spatial Scan Statistic and SaTScan Software
As in all medical research, it is important to determine whether any variation observed may reasonably be due to chance or not. This can be done using tests for spatial randomness, adjusting for the uneven geographical population density as well as for age and other known risk factors. One such test is the spatial scan statistic, which is used for the detection and evaluation of local clusters or hot-spot areas. This method is now in common use by various governmental health agencies including CDC and the state health departments in New York, Connecticut, Texas, Washington, Maryland, California and New Jersey (Appendix 1). Other test statistics are more global in nature, evaluating whether there is clustering in general throughout the map, without pinpointing the specific location of high or low incidence/mortality areas.

Using a space-time scan statistic, it has been possible to set up a system for regular time-periodic disease surveillance for the early detection of disease outbreaks. Such a system detects currently ‘active’ geographical clusters of disease, adjusting the statistical inference for the multitude of
possible geographical locations and sizes and cluster lengths considered and for the repeated
time-periodic analyses.

The spatial and space-time scan statistics have been implemented in two statistical software
packages. One of these is the freely available SaTScan software (Kulldorff et al 1998) that was
developed by and is distributed by the National Cancer Institute. The other is the ClusterSeer
software (BioMedware 2001), a commercial product.

B3. SaTScan Success Stories and Current Users

The circular spatial scan statistic and the accompanying SaTScan software has become widely
used by both governmental health departments and academic epidemiologists. A list of known
users of the spatial scan statistic and the SaTScan software is provided in Appendix 1. Most
applications have been for cancer and infectious diseases. It is also widely used in veterinary
epidemiology. The purpose is not only to find clusters when they exist, but also to dismiss
clusters that are likely chance occurrences. Some of the past and present applications include:

New York City Health Department: Daily surveillance for the early detection of disease
outbreaks. During the summer of 2001 it was successfully used for the early detection of dead
bird clusters to quickly detect local West Nile virus epicenters. Cluster findings led to preventive
measures such as targeted application of mosquito larvicide. During the spring of 2001 SaTScan
was successfully used as the early detection tool in a simulated bioterrorism exercise to train the
New York City mayor, his staff and health department officials in emergency preparedness and
conduct. Currently it is used for daily syndromic surveillance based on 911 emergency calls and
hospital emergency admissions. For additional information, see Mostashari et al (2001).

Washington State Health Department: Evaluation of a glioblastoma cluster alarm around
Seattle-Tacoma International Airport. Earlier analyses had been inconclusive as results depended
on geographical boundaries chosen to define this cancer cluster, and there were also questions
concerning pre-selection bias of airport area when testing the difference in the incidence rate
close versus further away from the airport. A SaTScan analysis for the county as a whole
revealed a non-significant cluster around the airport, adding weight to other evidence that it was
probably a chance occurrence. For additional information, see VanEenwyk et al (1999).

National Creutzfeldt-Jakob Disease Surveillance Unit and the Leicester Health Authority,
England: A very small but statistically significant ($p=0.004$) cluster with 5 cases of Creutzfeldt-
Jakob disease was found in Charnwood, Leicestershire, England. A detailed local
epidemiological investigation identified specific and unusual butcher shop practices as the likely
cause for this cluster. For additional information, see Bryant and Monk (2001), Cousens et al
(2001), and d’Aignaux et al (2002).

B4. General Theory of Scan Statistics

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 (1965ab), who looked at the problem in both one
the field, complementing an earlier edited volume (Glaz and Balakrishnan 1999).

In two or more dimensions, the events may be cases of leukemia, with an interest to see if there
are geographical clusters of the disease; they may be anti-personnel 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.

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. Depending on the application, different models are chosen, and depending on the model, the test statistic is evaluated either through explicit mathematical derivations and approximations or through Monte Carlo sampling (Dwass 1957). Due to inhomogeneous geographical population densities, there are no known asymptotic or approximate solutions for most disease surveillance problems, and Monte Carlo sampling is then used. Random data sets are generated under the known null hypothesis, and the value of the scan statistic is calculated for both the real data set and the simulated random data sets, and if the former is among the 5% highest, then the detected cluster is significant at the 0.05 level. While computer intensive, the Monte Carlo approach need not to be overly so, and it is possible to analyze data sets with 10,000+ geographical locations and 100,000 cases or more.


As for the scanning window, Naus (1965b), Loader (1991), Chen and Glaz (1996), Alm (1997, 1998) and Anderson and Titterington (1997) all considered rectangles. 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.

Based on the likelihood ratio test, Kulldorff (1997) presented a general mathematical model that includes all these cases, but even with the use of Monte Carlo sampling, it is not always computationally feasible to evaluate all possible window locations, sizes and shapes. While we no longer have to worry about the very difficult mathematics entailed in finding approximate or asymptotic solutions, we must now worry about efficient algorithms for evaluating a very large number of windows. The simplest shape to implement is the circle, and that is also the shape most commonly used, but it is also important to investigate more irregular shapes.

B5. Irregularly Shaped Clusters

The spatial extent of a disease cluster can reflect physical and man-made features of the environment such as rivers, highways, wind patterns, and topography (highlands versus lowlands). The circular scan statistic and its space-time generalization, which are designed to
detect compact clusters, provide poor delineation of irregularly shaped clusters resulting from such features (Figure 1). Irregularity of shape can also reduce the detection power of the circular scan statistic, or force it to report a single irregular cluster as a series of small clusters (Figure 2).

**Figure 1.** *Circular spatial scan statistic zonation (left) and cylindrical space-time scan statistic zonation (right).*

**Figure 2.** *Circular scan statistic represents a single actual cluster as a series of small clusters.*

**B6. Types of Geographical Disease Data**

For spatial health data, there are two different common probabilistic models, based on the Bernoulli and Poisson distributions respectively. With the Bernoulli model, there are cases and non-cases recorded as a binary (1/0) variable. These may represent people with or without a disease, or people with different types of disease. They may reflect cases and controls from a larger population, or they may together constitute the population as a whole.

With the Poisson model, the number of cases in each census area is assumed to be Poisson distributed. Under the null hypothesis, and when there are no covariates, the expected number of cases in each area is proportional to its population size, or to the person-years in that area. When there are covariates, such as age, gender, or socio-economic variables, the covariate adjusted expected number of cases is used. With either model, scan statistics must adjust for the uneven population density present in almost all populations, and the analysis is conditioned on the total number of cases observed.

For the Poisson model, it is necessary to have case and population counts for a set of census areas, as well as the geographical coordinates for each of those areas. For the Bernoulli model, the number of controls replaces the population counts. Separate census areas may be specified for individuals or data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc. To do a space-time analysis, it is also
necessary to have a time related to each case, and with the Bernoulli model, for each control as well.

A continuous response variable is a less common type of data in geographical public health surveillance. Examples include the study of geographical variation in blood pressure or melatonin levels.

C. Preparatory Studies

C1. Circular Spatial Scan Statistic


The spatial scan statistic was developed as a surveillance tool to detect and test the significance of local disease clusters, without making prior assumptions about their location or size, and adjusting the inference for the multiple testing inherent in the many potential sizes and locations considered. In mathematical terminology it is defined in terms of using a scanning window of any shape (Kulldorff 1997), but it has so far been implemented using a circular scanning window.

The method imposes a circular window on the map and lets its center move over the area so that at any given position, the window includes different sets of neighboring counties. If the window contains the centroid of a census area, then that whole census area is included in the window. For practical reasons, the center of the window is positioned only at the census tract centroids, and at each position, the radius of the circular window is varied continuously from zero up to a maximum radius so that the window never includes more than 50 percent of the total population. In this way, the circular window is flexible both in location and size. In total, the method creates a very large number of distinct circular windows, each with a different set of neighboring counties within it, and each a possible candidate for containing a cluster cancer. For each window, the method tests the null hypothesis against the alternative hypothesis that there is an elevated risk of breast cancer mortality within the window compared to outside.

For each circle, the observed and expected number of cases inside and outside the circle is noted, and the corresponding likelihood is calculated. This likelihood is maximized over all circles, identifying the window that constitutes the most likely disease cluster. The likelihood ratio for this window is calculated and constitutes the maximum likelihood ratio test statistic. Adjusting for multiple testing of multiple cluster locations and sizes, its distribution under the null-hypothesis and its corresponding $p$-value are obtained by repeating the same analytic exercise for 9999 random replications of the data set generated under the null hypothesis, in a Monte Carlo simulation (Dwass 1957). Calculations can be performed using the SaTScan software (SaTScan) developed at the National Cancer Institute specifically to implement the spatial scan statistic.
The spatial scan statistic has the following features, which make it particularly suitable for geographical cancer surveillance: (1) it adjusts both for the inhomogeneous population density and for any number of confounding variables such as age; (2) by searching for clusters without specifying their size or location the method ameliorates the problem of preselection bias; (3) the likelihood ratio based test statistic takes multiple testing into account, and delivers a single \( p \)-value for the test of the null hypothesis; and (4) if the null hypothesis is rejected, we can specify the approximate location of the cluster that caused the rejection.

Figure 3. The most likely cluster as found by the circular spatial scan statistic (left) and the elliptic spatial scan statistic (right) for an analysis of county-based breast cancer mortality in Northeastern United States, 1988–1992. For the circular-detected cluster, the relative risk is 1.07 and \( p=0.0001 \) (Kulldorff et al, 1997), while the elliptic-detected cluster has a relative risk of 1.08 and \( p=0.0001 \). Note that the elliptic-detected cluster is not connected, since the New York City area is not part of the cluster.

C2. Early Detection of Disease Outbreaks


One of the most important problems in geographical disease surveillance is the early detection of disease outbreaks, as even only a few days advanced warning will give public health officials a head start on implementing disease prevention and control measures. This is especially important for infectious diseases. This was accomplished using a cylindrical space-time scan statistic, where the circular base represents the geographical area and the height of the cylinder represents time. For each circular area, the method scans many different time intervals of different length in order to minimize pre-selection bias. The inference is adjusted for all the different spatial areas and time intervals considered, as well as for the many time-periodic analyses that are performed in a typical surveillance setting.
During the summer of 2002, the method was applied to West Nile virus surveillance in New York City, and it is currently applied on a daily basis by the New York City health Department for syndromic surveillance using hospital admissions and 911 emergency call data.

C3. Elliptic Spatial Scan Statistic

Using exactly the same mathematical framework as for the circular scan statistic, the principal change needed for the elliptic spatial scan statistic is that it evaluates a number of additional elliptic areas of different size, shape and angle as potential clusters. The shape is defined as the ratio of the length and width of the ellipse, and the angle is the direction of the longest axis. An ellipse with a shape of 1 is a circle, which is therefore included as a special case. We investigated different sets of shapes and angles to find an appropriate balance between optimal cluster detection and computing time. For the breast cancer mortality data, Figure 3 shows the most likely elliptic cluster, which has a shape of 4, an angle of 30 degrees, a relative risk of 1.08, and a \( p \)-value of 0.0001.

C4. Echelons, Remote Sensing, and Tree-Structured Scan Statistic


Ecosystem health assessment involves geospatial landscape pattern analysis (Johnson et al 2001, 2001ab; Patil 2001; Patil and Taillie 1999, 2002) thematic map and change map analysis (Patil and Taillie 2001, 2002, 2003), cellular surface analysis (Myers and Patil 2002; Myers et al 1997, 1999; Johnson et al 1998), critical area detection and delineation (Patil et al 2000, 2002ab, 2003ab; Patil 2002), modeling and simulation (Patil and Taillie 1999, 2001; Patil et al 2000). In the recent past, much work has been done at Penn State for identifying the highs and lows and connecting corridors in the spatial distribution of environmental variables. This work has resulted in a methodology called echelons (Myers and Patil 2002). The current proposal has been largely motivated by a desire to merge the methods of echelon analysis with those of the spatial scan statistic to help improve on both.
Echelons synthesize the topological structure of spatially mapped environmental indicators for objective analyses of complex hierarchies of spatial variation across landscapes. The environmental indicator is regarded as a surface variable that represents a virtual topography as depicted (in one-dimensional profile) in Figure 4. Echelons are structural entities consisting of peaks, foundations of peaks, foundations of foundations, and so on in an organizational hierarchy. It is natural to display the echelon hierarchy as a tree and, in this form, echelons have proven effective for detecting and analyzing patterns of concentration and connectivity for biodiversity, landscape change, urban sprawl, etc. (Myers et al 1995, 1997; Johnson et al 1998; Myers et al 1999; Kurihara et al 2000; Smits and Myers 2000).

Figure 4. *Echelon decomposition of a surface (left) and associated echelon tree (right).*

Contemporary study of human disease as a component of ecosystem health entails a spatial scan statistic (Kulldorff and Nagarwalla 1995) for detecting geographic clusters of disease and other responses that are significantly elevated with respect to the regional setting. In conjunction with the spatial scan statistic, echelon analysis can more clearly delineate the cluster boundaries for focus of investigation as depicted in Figure 1.

Consider a tessellation of a geographic region with a response value assigned to each cell of the tessellation. Echelon analysis is an objective means of analyzing the spatial structure of the response function. In environmental and remote sensing applications, the tessellation is often regular (for example, pixels in an image or EMAP hexagons) but the basic logic of the method does not require this. Examples of environmental response functions include: (i) gray-scale level in a raster image, (ii) a pixel-referenced index of change between two images, (iii) abundance of a species across a cell, and (iv) species richness across a cell.

There is an analogy between echelon analysis and Morse theory. The latter considers smooth response functions $G(x)$ defined over a differentiable manifold. For each value $g$ of the response function, the “upper level set” is defined as $\{ x : G(x) \geq g \}$. Morse (1934) studied how the topological structure (specifically, the higher order connectivity) of the upper level sets changes as the level $g$ decreases; see Milnor (1963) and Matsumoto (2002).

In echelon analysis, the response function $G$ is defined over a finite set $M$ of cells instead of a differentiable manifold. The differentiable structure on $M$ is replaced by a nearest-neighbor structure. For example, a raster image has two standard nearest-neighbor structures depending
on whether diagonally adjacent cells (pixels) are considered to be neighbors. A “path” in $M$ is then a sequence of cells in $M$ with the property that every two consecutive cells in the sequence are nearest neighbors. A subset $Z$ of $M$ is “connected” if any two cells in $Z$ can be joined by path that lies entirely within $Z$. Every subset $X$ of $M$ can be written uniquely as a disjoint union of maximal connected subsets called the connected components of $X$. The term “connectivity” of $X$ refers to the number and identity of the connected components $Z$ of $X$.

Echelon analysis studies how the connectivity of the upper level sets $\{x : G(x) \geq g\}$ changes as the level $g$ decreases. For example, if the response is species richness in a cell then for large values of $g$ the upper level set may consist of a few isolated pockets of high species richness. As $g$ decreases, corridors of lower richness may join up these pockets. This changing pattern of connectivity can be quite varied and complex. At the simplest, there might be only one primary region of high richness that grows by gradually accumulating neighboring cells as $g$ decreases. A slightly more complex pattern has two primary regions, each growing by gradually accumulating their neighboring cells, but which do not join up until $g$ is very small. Echelon analysis represents each of these various patterns as a rooted tree with a height function defined on the nodes of the tree. Nodes at height $g$ are in one-to-one correspondence with the connected components of the upper level set $\{x : G(x) \geq g\}$.

Echelon analysis may be used in conjunction with the spatial scan statistic to more clearly delineate cluster boundaries, since echelon families identify the spatial connectivity of a response surface. For example, two isolated first order echelons may be connected by a common second order echelon, as identified by “saddle point” mapping units. Echelons at any hierarchical level may be tested for statistical significance by the spatial scan statistic approach. Therefore, the combination of these two different methods will result in the determination of spatially disjoint areas of significantly elevated disease rates. Essentially, echelon analysis mechanizes and objectifies the way a person may look at a map and quickly determine a reasonable set of candidate zones, while eliminating many other zones as obviously uninteresting.

Therefore, echelon analysis used in conjunction with the spatial scan statistic may improve disease surveillance for programs that currently apply the scan statistic. For example, we have revisited the North Carolina SIDS mortality data over counties during 1974–84 analyzed by Kulldorff (1997). The following diagrams show the hotspot results of the two approaches based on circles and echelons as the choices for candidate zones.

**Figure 5.** Primary and secondary clusters for North Carolina SIDS (county-wide). Circle-based clusters are shown on the left and echelon-based clusters on the right. Primary cluster is the same for both search methods.
The echelon-based scan statistic can be improved further by considering the entirety of nodal connected components consisting of cells cumulated as of each node. This has led us to the proposed tree-structured scan statistic discussed in Section D1c of the proposal. The tree described in Section D1c has a node for every connected component of every upper level set; the echelon tree is the subtree consisting of only the leaf nodes and the junction nodes.

**C5. Simulated Annealing Scan Statistic**


This paper proposes a new graph-based strategy for the detection of spatial clusters of arbitrary geometric form in a map of geo-referenced populations and cases. Our test statistic is based on the likelihood ratio test previously formulated by Kulldorff and Nagarwalla (1995) for circular clusters and in Kulldorff (1997) for general zones. Scan methods usually are restricted to the search of clusters of fixed geometric form, like circles or rectangles, instead of arbitrarily shaped connected clusters. The reason for this is that the task of analyzing all possible clusters is computationally infeasible. The simulated annealing approach commonly used in physics (see e.g., Aarts and Korst 1989; Winkler 1995) gives a convenient answer to this difficulty by analyzing only the most interesting cluster candidates and finding a quasi-optimal solution. A new technique of adaptive simulated annealing was developed, focused on the problem of finding the local maxima of the likelihood function over the space of the connected subgraphs of the graph associated to the regions of interest. Given a map with \( n \) regions, on average this algorithm finds a quasi-optimal solution after analyzing \( s n \log(n) \) subgraphs, where \( s \) depends on the cases density uniformity in the map. It was implemented in a conveniently fast C computer language code. Although the computational effort is greater compared to other scan methods, experiments show that the simulated annealing method identifies clusters and tests their statistical significance for real life problems in a reasonable amount of time using modest computer resources.

*Figure 6. The clusters found in Belo Horizonte metropolitan area. At left, the cluster found by the circular scan statistic, and to the right the one found by the simulated annealing scan statistic. The light gray areas have zero cases.*
We used our method to study the geographical distribution of homicides in Belo Horizonte, a city of two million inhabitants in Southeast Brazil, as shown in Figure 6. For the simulated annealing method, the most likely cluster is placed along a large expressway, which is a degraded urban area. Hence, this cluster departs substantially from the circular shaped clusters and, if real, it would be hard to detect using the circular spatial scan statistic. We also note that it has fewer cells, a smaller population, more cases, and as a consequence a higher case density. The likelihood ratio for this cluster is substantially higher. We used 999 random allocations in order to test cluster significance, and found $p$-values of 0.001 for both clusters. Using a 1.8 GHz PC computer, the simulated annealing analysis was done in 6 minutes.

C6. Benchmark Spatial Data for Power Evaluations


In this paper we present a collection of 1,220,000 simulated benchmark data sets generated under 51 different cluster models and the null hypothesis, to be used for power evaluations. We then used these to estimate and compare the power of the circular spatial scan statistic (Kulldorff 1997), Tango’s Maximed Excess Events Tests (2000), and Bonnetti-Pagano’s $M$ statistic (2002). By making the simulated data sets publicly available, new tests can easily be compared with previously evaluated tests by analyzing the same benchmark data.

For the population data, we used the 1990 women population in 245 counties in the Northeastern United States (Kulldorff et al 1997). Two different types of clustering models were used for the alternative hypothesis. One was a hot-spot cluster alternative where the risk of disease is higher in one or more cluster areas, but where the locations of cases are independent of each other. The other was a global clustering alternative, where each county has the same number of expected cases under the null and alternative models, but where the locations of cases are dependent on each other.

D. Research Design and Methods

D1. Different Methodological Approaches

D1a. Spatial Scan Statistic and Search Strategies

The spatial scan statistic deals with the following situation. A region $R$ of Euclidian space is tessellated or subdivided into cells that will be labeled by the symbol $a$. Data is available in the form of a count $Y_a$ (non-negative integer) on each cell $a$. In addition, a “size” value $A_a$ is associated with each cell $a$. The cell sizes $A_a$ are regarded as known and fixed, while the cell counts $Y_a$ are independent random variables. Two distributional settings are commonly studied:

1) **Binomial**: $A_a = N_a$ is a positive integer and $Y_a \sim \text{Binomial} \ (N_a, p_a)$, where $p_a$ is an unknown parameter attached to cell $a$ with $0 < p_a < 1$.

2) **Poisson**: $A_a$ is a positive real number and $Y_a \sim \text{Poisson} \ (\lambda_a A_a)$, where $\lambda_a > 0$ is an unknown parameter attached to cell $a$. 
Each distributional model has a simple interpretation. For the binomial, a collection of $N_a$ items resides in cell $a$ and each item has a certain trait independently with probability $p_a$. The cell count $Y_a$ is the number of items having the trait. For the Poisson, $A_a$ is the size (perhaps hypervolume) of the cell $a$, and $Y_a$ is the realized number of points in a Poisson process of intensity $\lambda_a$ across the cell.

The spatial scan statistic seeks to identify “hotspots” or “clusters” of cells that have an elevated response compared with the rest of the region. However, by elevated response we do not mean large values for the raw counts $Y_a$; instead, we mean large values for the rates,

$$G_a = Y_a / A_a.$$ 

In other words, the cell counts are adjusted for cell sizes before comparing cell responses. The scan statistic easily accommodates other rate adjustments, such as for age or for gender. In some situations, one also wishes to find regions of low response; this is easily accomplished by reversing the direction of the rate function.

A collection of cells from the tessellation should satisfy several geometrical properties before it could be considered as a candidate for a hotspot cluster. First, the union of the cells should comprise a connected subset of the region $R$. All such collections of cells will be referred to as zones and the set of all zones is denoted by $\Omega$. Thus, a zone $Z \in \Omega$ is any collection of cells that are connected. Second, the zone should not be excessively large—for, otherwise, the zone instead of its exterior would constitute background. This restriction is generally achieved by limiting the search for hotspots to zones that do not comprise more than, say, fifty percent of the population.

![Figure 7. A tessellated region. The collection of shaded cells in the left-hand diagram is connected and, therefore, constitutes a zone in $\Omega$. The collection on the right is not connected.](image)

The following hypothesis testing scenario formalizes the cluster detection and delineation problem. For definiteness, we use the binomial distributional setting.

$$H_0: p_a \text{ is the same for all cells in region } R, \text{ i.e., there is no cluster.}$$

$$H_1: \text{ There is a non-empty zone } Z \text{ (connected union of cells) and parameter values}$$

$$0 < p_0, p_1 < 1 \text{ such that}$$

$$p_a = \begin{cases} 
  p_1 & \text{for all cells } a \text{ in } Z \\
  p_0 & \text{for all cells } a \text{ in } R - Z
\end{cases} \quad \text{and} \quad p_1 > p_0.$$
The zone $Z$ specified in $H_1$ is an unknown parameter of the model. The full model, $H_0 \cup H_1$, involves three unknown parameters:

$$Z, p_0, p_1 \text{ with } Z \in \Omega \text{ and } p_0 \leq p_1.$$ 

The null model, $H_0$, is the limit as $p_1 \to p_0$; however, the parameter $Z$ is not identifiable in the limit. For given $Z$, the likelihood estimates of $p_0$ and $p_1$ can be written down explicitly which allows us to determine the profile likelihood for $Z$:

$$L(Z) = \max_{p_0, p_1} L(Z, p_0, p_1) = L(Z, \hat{p}_0, \hat{p}_1).$$

The difficult part of hotspot estimation lies in maximizing $L(Z)$ as $Z$ varies over the collection $\Omega$ of all possible zones. In fact, $\Omega$ is a finite set but it is generally so large that maximizing $L(Z)$ by exhaustive search is impractical. Two different search strategies are available for obtaining an approximate solution of this maximization problem:

1) **Parameter-space reduction.** Replace the full parameter space by a subspace $\Omega_0 \subset \Omega$ of a more manageable size. The profile likelihood $L(Z)$ is then maximized by *exhaustive search* for $Z \in \Omega_0$. This works well if $\Omega_0$ contains the MLE for the full $\Omega$ or at least a close approximation to that MLE. Parameter space reduction is roughly analogous to doing a grid search in conventional optimization problems where the parameter space is a region in Euclidian space.

2) **Stochastic optimization methods.** These methods include genetic algorithms (Knjazew 2002) and simulated annealing (Aarts and Korst 1989; Winkler 1995). These are iterative procedures that converge, under certain assumptions, to the global optimum in the limit of infinitely many iterations. In practice, the procedure must be terminated after only finitely many iterations giving only an approximate solution. A more serious limitation is that these procedures are computationally intensive enough that they can be difficult to replicate many times in a simulation study to determine null distributions.

The traditional (circular) scan statistic uses expanding circles to determine a reduced list $\Omega_0$ of candidate zones $Z$. By their very construction, these candidate zones tend to be compact (circular) in shape and may do a poor job of delineating the actual clusters. The proposed research will develop, evaluate, and compare three strategies for identifying clusters of arbitrary shape. These are referred to as (i) the ellipse spatial scan statistic, (ii) the tree-structured scan statistic, and (iii) the simulated annealing scan statistic. The first two of these strategies employ parameter-space reduction while the third employs stochastic optimization.

Similar in spirit to the circular scan statistic, the elliptic scan statistic identifies its reduced parameter space $\Omega_0 = \Omega_{\text{ellips}}$ via ellipses having a range of centers, sizes, eccentricities, and orientations. With so many variable features, the cardinality of $\Omega_{\text{ellips}}$ can be rather large with correspondingly heavy demands on computer time.

Both the circular and elliptic scan statistics have reduced parameter spaces that are determined entirely by the geometry of the tessellation and do not involve the data in any way. The tree-structured scan statistic takes an adaptive point of view in which $\Omega_0$ depends very much upon
the data. In essence, the adjusted rates define a piece-wise constant surface over the tessellation, and the reduced parameter space \( \Omega_{ULS} \) consists of all connected components of all upper level sets of this surface. The cardinality of \( \Omega_{ULS} \) does not exceed the number of cells in the tessellation. Furthermore, \( \Omega_{ULS} \) has the structure of a tree (under set inclusion), which is useful for visualization purposes and for expressing uncertainty of cluster determination in the form of a hotspot confidence set on the tree. Since \( \Omega_{ULS} \) is data-dependent, this reduced parameter space must be recomputed for each replicate data set in a simulation (in determining null distributions for example).

The simulated annealing scan statistic examines a statistical sample of zones from the full zonal space \( \Omega \). The number of zones in the sample is typically between 10,000 and several hundred thousand. Two zones in \( \Omega \) are said to be neighbors if they differ by exactly one cell. The sample of zones is generated by repeatedly jumping from the current zone to one of its neighboring zones according to a complex set of stochastic transition rules and stopping rules. A number of annealing runs are made. For the first run, the starting zone is obtained by applying the circular scan statistic. On the remaining runs, the starting zone is a randomly selected cell.

Although spatial scan statistics are commonly applied to tessellated data, both the annealing and tree-structured approaches have an abstract network or graph (i.e., nodes and edges) as their starting point. Accordingly, these two approaches can also be applied to data defined over a network, such as a subway, water or highway systems. In the case of a tessellation, the abstract graph is obtained by taking its nodes to be the cells in the tessellation. Two cells are joined by an edge if they are adjacent in the tessellation. Both approaches are flexible regarding the definition of adjacency. For example, one may declare two cells as adjacent if (i) their boundaries have at least one point in common, or (ii) their common boundary has positive length, or (iii) the length of their common boundary exceeds some specified percentage of their total perimeter. The user is free to adopt whatever definition of adjacency seems most appropriate to the problem at hand.

**D1b. Elliptic Spatial Scan Statistic**

In its simple straightforward version, the elliptic spatial scan statistic has already been developed and tested, and it is shown to be able to detect and evaluate the significance of non-circular clusters. Hence, we are in a sense already prepared to move forward and compare the elliptic scan statistic with the simulated annealing and tree structured scan statistics as soon as the development of these methods is completed.

There are a few issues that need further developmental investigation though. Since there are many more possible angles and shapes for the elliptic window than for a circular window, we may be more likely to find an elliptic most likely cluster even when the true cluster is circular or when null hypothesis is true. This is in a sense similar to regression, where the fit is better with more variables even if the additional variables have no explanatory power. What we would like to do, is to adjust the analysis for the fact that many more elliptic than circular clusters are considered, and to take it one step further, for the fact that there are more narrow elliptic clusters than fat ones.

This will be done as follows. Under the null hypothesis, we can for each simulated data set obtain the most likely cluster for each elliptic shape, including the circle that corresponds to a shape of 1. The shape is as previously described defined as the ratio of the length to the width, or in mathematical language, the ratio of the major axis to the minor axis. Over all the simulated
data sets, we can then calculate the average of the maximum log likelihoods over the circles as well as over each of the elliptic shapes. The difference in these averages is then what we will expect the differences that is smaller indicates evidence that the circular cluster is the best fit, while if the difference is larger than an elliptic cluster fits the data better. In selecting and ranking the clusters, the log likelihood for each elliptic cluster shape will be modified to reflect its average difference compare to the circular. We will then investigate how this alternative approach compares to the simple straightforward version in terms of power and cluster accuracy for a wide variety of alternative cluster models.

**D1c. Tree-Structured Scan Statistic**

The tree-structured scan statistic is an adaptive approach in which the reduced parameter space \( \Omega_0 = \Omega_{ULS} \) is determined from the data by using the empirical cell rates

\[
G_a = Y_a / A_a.
\]

These rates determine a function \( a \rightarrow G_a \) defined over the cells in the tessellation. This function has only finitely many values (called levels) and each level \( g \) determines an upper level set (ULS), \( U_g \), defined by

\[
U_g = \{ a : G_a \geq g \}.
\]

Figure 8 depicts some upper level sets in the case of a smoothly varying rate function \( G \). The horizontal axis represents the spatial region under study, which is depicted as one-dimensional for display purposes. For application to the scan statistic, \( G \) is piecewise constant and each upper level set \( U_g \) is a finite collection of cells from the tessellation. As shown in Figure 8, upper level sets do not have to be connected. Figure 8 also shows how the connectivity of the ULS can change as the level drops from \( g \) to \( g' \). The ULS reduced list of candidate zones, \( \Omega_{ULS} \), consists of all connected components of all possible upper level sets.

**Figure 8.** Schematic (smooth version) for the intensity surface, shown in profile. The shaded portion of the horizontal axis on the left is the upper level (ULS) at intensity level \( g \). Each of the
three pieces \( Z_1, Z_2, Z_3 \) is a connected component of the ULS and, therefore, one of the candidate zones in \( \Omega_{ULS} \). As the level drops from \( g \) to \( g' \), the connectivity of the upper level set changes. The diagram on the right illustrates the three types of change: (i) zones \( Z_1 \) and \( Z_2 \) grow in size and eventually coalesce into a single zone \( Z_4 \), (ii) Zone \( Z_3 \) simply grows to \( Z_5 \), and (iii) Zone \( Z_6 \) is newly emergent.

A tree structure can be defined on the reduced parameter space \( \Omega_{ULS} \). The nodes of the tree are the members of \( \Omega_{ULS} \), i.e., the candidate zones. Two nodes \( Z, Z' \in \Omega_{ULS} \) are joined by an edge if

(i) \( Z \) is a proper subset of \( Z' \), written as \( Z \subsetneq Z' \) and

(ii) there is no node \( W \in \Omega_{ULS} \) such that \( Z \subsetneq W \subsetneq Z' \).

Figure 9 shows the tree structure for the intensity surface displayed in Figure 8.

![Intensity Surface Diagram](image)

**Figure 9.** ULS connectivity tree for piece-wise constant version of the intensity surface displayed in Figure 8. The four leaf nodes correspond to modes (peaks) of the intensity surface. The root node represents the entire region. Junction nodes (A, B and C) occur when two (or more) connected components coalesce into a single connected component. The collection of all nodes in the tree is the reduced zonal parameter space \( \Omega_{ULS} \). A profile likelihood value \( L(Z) \) can be attached to each of these nodes. The node \( \hat{Z} \) that maximizes \( L(Z) \) across the tree is the ML estimate for the hotspot cluster under the tree-structured scan statistic.

A consequence of adaptivity of the ULS approach is that the reduced parameter space \( \Omega_{ULS} \) depends on the data \( \{Y_a\} \) and is not a feature of the tessellation alone. Accordingly, \( \Omega_{ULS} \) must be recalculated for each replicate in a simulation study. Part of the research effort will be the development of efficient algorithms for this calculation. Cell adjacency determines a symmetric relation on the set of all cells. Finding the connected components for an upper level set is
essentially the issue of determining the transitive closure of the adjacency relation on the cells in
the upper level set. Several generic algorithms are available in the computer science literature
(Cormen et al, 2001, Section 22.3 for depth first search; Knuth, 1973, p. 353 or Press et al, 1992,
Section 8.6 for transitive closure). But special features of the ULS connectivity problem permit
enhanced efficiency. We represent cell adjacency by a zero-one adjacency matrix \( A \) whose rows
and columns are labeled with the cells of the tessellation. Entry \( A_{ab} \) equals 1 if cells \( a \) and \( b \)
are the same cell or are adjacent in the tessellation. Otherwise, \( A_{ab} \) vanishes. The cells (row
and column labels) are arranged in order of decreasing intensity \( Z_a \) so that the
adjacency matrix for any upper level set is a square submatrix in the northwest corner of the full
adjacency matrix \( A \). This reordering of the rows and columns of \( A \) is the only data dependent
part of the algorithm. As the level drops cells are added one after another and one has to
determine how the connectivity changes with each addition of a cell. As shown in Figure 8,
there are three possibilities:

(i) Two or more connected components coalesce into one. This occurs when the new cell
is adjacent to several existing connected components and forms a bridge among them.

(ii) An existing connected component grows in size. This occurs when the new cell is
adjacent to exactly one existing connected component.

(iii) A new connected component is formed. This occurs when the newly added cell is not
adjacent to any of the existing connected components.

Execution time will depend on the number of nodes in the tree. However, as we trace down the
tree from leaf nodes to root, each cell of the tessellation makes its first appearance in a uniquely
determined node. This implies that the number of nodes in the tree is less than or equal to the
number of cells in the tessellation. Equality holds when distinct cells \( a \) have distinct intensity
levels \( Z_a = Y_a / A_a \). Computer efficiency can be further improved since it is not necessary to
compute the portion of the tree below a specified level (e.g., mean or median intensity) since the
Corresponding cells are background and not plausible locations for hotspot clusters.

Confidence Sets for Hotspot Estimation

The hotspot MLE is just that—an estimate. Removing some cells from the MLE and replacing
them with other cells can generate an estimate that is almost as plausible in the likelihood sense.
This zonal estimation uncertainty can be expressed by a confidence set of zones. For example, if
we wish to determine if a particular cell (e.g., county, zip code) belongs to the hotspot, it would
not be appropriate to ask if the cell belongs to the zonal MLE \( \hat{Z} \). It would be better to ask if the
cell belongs to at least one of the zones in a confidence set for the hotspot.

Confidence Set Determination

We employ the standard duality between confidence sets and hypothesis testing, namely that the
confidence set consists of all null hypotheses that cannot be rejected at a specified significance
level \( \alpha \) (Bickel and Doksum 1977, p. 179; Lehmann 1986, Section 3.5, Theorem 4). The
confidence level is then \( c = 1 - \alpha \). Alternatively, the confidence set contains all null hypotheses
for which the \( p \)-value exceeds \( 1 - c \).
We need to formulate the null hypotheses for the present setting in which the parameter space is \( \Omega_{uls} \): the set of all zones that are connected components of upper level sets of the intensity function. The confidence set will be a subset of \( \Omega_{uls} \) and a particular zone \( Z_0 \in \Omega_{uls} \) is in the confidence set if we cannot reject the following null hypothesis:

\[
\bar{H}_0: \text{There are binomial parameters } p_i \geq p_o \text{ such that } \\
p_a = \begin{cases} 
p_i & \text{for all cells } a \text{ in } Z_0 \\
p_o & \text{for all cells } a \text{ outside } Z_0 
\end{cases}
\]

This hypothesis is to be tested against the general alternative in which the hotspot zone \( Z \) is allowed to vary freely over \( \Omega_{uls} \). Schematically, then, we are testing \( \bar{H}_0 : Z = Z_0 \) versus \( \bar{H}_1 : Z \neq Z_0 \) and the confidence set consists of all zones \( Z_0 \) for which \( \bar{H}_0 \) cannot be rejected.

We carry out the test using the likelihood ratio statistic, LR, for which two questions need to be addressed:

1. How, as a technical matter, is the null distribution to be simulated for given \( Z_0 \)?
2. How do we handle and interpret multimodality of the LR statistic, giving rise to “disconnected” confidence sets?

For the first question, we note that the null hypothesis \( \bar{H}_0 \) involves, as nuisance parameters, the rates \( p_o \) and \( p_i \) outside and inside zone \( Z_0 \). Conditioning on the totals outside and inside can eliminate these nuisance parameters so that the simulation amounts to sampling without replacement outside and inside this zone. Dependence of the null distribution upon \( Z_0 \) is a matter that needs to be examined.

The second question is nicely addressed by the tree structure on our reduced parameter space \( \Omega_{uls} \). The nodes of \( \Omega_{uls} \) are our candidate zones and a likelihood-based confidence set is an upper level set of the likelihood ratio function defined over the tree (Figure 10). As shown in the Figure 10, this upper level set may have several connected components, exactly one of which contains the MLE. This is because we cannot say with statistical certainty that the MLE correctly identifies the hotspot locus. The other connected components are plausible (at the current confidence level) alternative loci. The nodes comprising a connected component are the plausible delineations of that hotspot locus. The connectivity and the makeup of the connected components change with the confidence level, corresponding to varying degrees of plausibility.
Figure 10. An upper level set of the LR function is shown on the ULS tree. This upper level set is a confidence set for the hotspot locus and, in this case, has two connected components. The different connected components correspond to different hotspot loci while the nodes within a connected component correspond to different delineations of that hotspot – all at the appropriate confidence level. Extremity nodes are nodes within the confidence set but with a neighboring node outside the confidence set. Simulation at extremity nodes checks null distribution constancy and estimates confidence levels. As the confidence level increases, the confidence region grows in size and its connectivity may change. Connectivity changes are of these types: (i) several components coalesce; multiple hotspot loci merge into a single locus with a complex internal structure. (ii) a connected component grows in size giving a larger set of plausible hotspot delineations. (iii) a new connected component emerges corresponding to an additional hotspot locus mandated by the higher confidence level.

D1d. Simulated Annealing Scan Statistic

A major problem with scan statistics methods is the fixed shape of the clusters to be detected. The reason for such a restriction is the computationally infeasible number of possible areas to be tested. However, in real situations, we frequently find spatial clusters with quite different shapes from circular ones. Increased risks along rivers, transport ways or power lines create clusters with shapes highly different from circular ones. Usually the environmental or social causes may not have a circular symmetry. The graph-based framework used in the simulated annealing method discussed in section C5 has a potential to overcome these problems, but it may be improved in several ways, as follows.

Oriented Graphs

The graphs that we are working with until now are non-oriented ones, i.e., if A is neighbor of B, then B is also a neighbor of A. In certain situations it may be useful to consider oriented graphs, that is, A is neighbor of B does not necessarily implies that B is a neighbor of A. Such situations
arise in the investigation of contamination in pipelines, where the water flows in just one
direction. Another important example appears when we need to track a disease spread along
metropolitan subways.

We propose an adjacency matrix \( [a_{ij}] \) with entries \( a_{ij} = 1 \) if node \( i \) is neighbor of node \( j \), and \( a_{ij} = 0 \) otherwise. With this definition, we can even connect census regions that are not
neighbors in the planar map, but are important end points for the traffic flux of people, as for
example between residential and working areas of a city. In fact, the flux of people in this case is
even more significant than the flux between adjacent residential census regions. In this situation
the adjacency matrix can be modified to allow non-negative real numbers as a measure of the
people’s flux between the census regions. In the context of fast propagation diseases, we can thus
work with more realistic maps.

**Cohesion of a Zone**

When the census regions in the map do not have similar sizes, a problem may occur because the
length of the edges joining the nodes may vary widely. For instance, if a zone has a large census
region amidst other very small ones, then we cannot expect that that this zone is as strongly
connected as another zone with only small census regions. So, in order to penalize the zones with
larger regions we modify the likelihood ratio as follows. Given the collection of all edges in a
connected subgraph \( G \) with \( n \) nodes (\( n > 1 \)) in decreasing order of size, we define its cohesion
\( C(G) \) as a function of the minimal number of edges that are necessary to remove from \( G \) in
order to disconnect it. So, if \( a_1 \geq a_2 \geq \cdots \geq a_m \) are the lengths of the \( m \) largest edges and the
subgraph becomes disconnected after removing these \( m \) edges (but not if we remove only the
first \( m-1 \) edges), then \( C(G) = \sum_{k=1}^{m} a_k^{-1} \). So, instead of using the likelihood \( L(G) \) we use
\( C(G) \cdot L(G) \). We also define the cohesion of a 1-node subgraph \( \{ x \} \) as the average cohesion of
all the connected subgraphs with 2 nodes that contain \( x \). There is a smart way of computing
\( C(G) \), without removing the \( m \) edges one by one and then testing for the connectedness of the
remaining graphs. We use a bisection procedure to remove the edges, so instead of making \( m \)
tests for connectedness we use at most only \( \log_2 n \) tests. By this way, the presence of a very
large census region within a zone decreases its cohesion, and as a consequence decreases its
likelihood. With this new definition, a more realistic analysis could be done, because we take
into account not only the topology of the graph, but also the distance between the nodes.

**Compactness of a Connected Zone**

In some situations, we would like to penalize the likelihood of cluster candidates that are too
much different in shape from a circular area. This can be useful if we want to avoid ill-shaped
clusters that cannot be ascribed to any relevant specific cause known in advance. For this
purpose, we define the compactness of a connected zone \( z \) as

\[
K(z) = \frac{4\pi A(z)}{H(z)^2},
\]
where $A(z)$ is the area of $z$ and $H(z)$ is the perimeter of the convex hull of $z$. So a disk is a zone with maximum compactness, namely 1. The area of each zone $z$ is easy to compute as a sum of the areas of the individual census regions that are generally available. But we need to compute an approximation for $H(z)$ from the elements of the graph. Our approach is as follows. We first compute $P_0(z)$, the perimeter of the convex hull of the nodes that forms the zone $z$ and then $P_1(z)$, the perimeter of convex hull of the nodes that are neighbors of the zone $z$. $H(z)$ is approximated as the average of $P_0(z)$ and $P_1(z)$. In order to compute the convex hull of $z$ we use the Quick Hull algorithm (see, e.g., Berg et al 2000), which is generally efficient for random distributed points in the plane. For the computation of the convex hull of the neighbors of $z$ we may use the Graham’s scan algorithm (Berg et al 2000), which is best suited for a set of points in the plane in a ring-like disposition.

### D2. General Conceptual Issues

#### D2a. Data Aggregation and Its Effect on Performance

Data aggregation averages case occurrences across entire cells, reducing the ability to detect and accurately map disease clusters. Sensitivity of the proposed scan statistics to aggregation will be assessed and compared by aggregating existing data sets to coarser levels and evaluating the effects on power and accuracy of cluster estimation.

#### D2b. Tessellation Geometry and Its Effect on Performance

Performance of the proposed scan statistics is affected by tessellation geometry (variation of cell size and cell shape) as well as aggregation level (average cell size). A single actual cluster that is bisected by a large cell may appear either as not significant or as two distinct clusters depending on whether the large cell is included in the candidate zone (Figure 11). The tree-structured and annealing scan statistics can deal with such artifacts of tessellation geometry by declaring cells as “adjacent” even when they are not physically contiguous.

![Figure 11. Actual cluster (shaded) appears as two clusters because a large cell of comparatively low overall incidence rate bisects it. Only pertinent cells are shown.](image)

#### D2c. Parametric Approximations for Null Distributions

Hypothesis testing for the scan statistic employs the likelihood ratio (LR) test. Under “standard conditions,” LR null distributions are asymptotically chi-squared with degrees of freedom equal to the codimension of the null model inside the full model. Parameter structure in the scan statistic setting is non-standard, however, since $(i)$ the zonal space ($\Omega$ or $\Omega_0$) is finite discrete
and (ii) the zonal parameter \( Z \) is not identifiable under the null hypothesis (Davies 1977). Null distributions are, therefore, obtained by Monte Carlo simulation for each application of the scan statistic. Data simulation is the same for the three approaches proposed here, but each approach must apply its own separate data handling to each replicate data set. For example, the tree-structured scan statistic must calculate a new tree for each replicate, while the annealing scan statistic must run separate searches on each replicate. About 1,000 to 10,000 replicates are desirable for accurate determination of \( \alpha = 0.05 \) critical points and consume most of the computer time needed for application of the scan statistic method.

We will investigate whether the simulated null distributions can be accurately approximated (especially in the upper tails) across a wide range of conditions by standard parametric families of probability distributions. Potential families include the chi-squared, the gamma (scaled chi-squared distribution), and the beta of the second kind (scaled \( F \)-distribution). Assuming a good-fitting family can be identified, the parameter values will depend upon numerous conditions such as aggregation level, tessellation geometry, and population sizes and their spatial distribution across the tessellation. Parameter values will also depend upon which of the three scan statistic approaches is used. For example, Duczmal L, Assunção RA (2002) simulated the null distributions for Belo Horizonte (Section C5) and computed the mean values of \(-2/(LR)\) across 999 replicates. These mean values, corresponding to the degrees of freedom for approximating chi-squared distributions, were 10.3 for the circular scan statistic and 39.8 for the simulated annealing scan statistic. No general \textit{a priori} rules relating parameter values to these conditions can be expected, so parameters for approximating null distributions will have to be estimated using simulated data. But, this should reduce substantially the number of replicates required and should also allow extrapolation to smaller significance values and smaller \( p \)-values. In addition, fitted null distributions and their parameters are of independent interest for characterizing and contrasting different geographical regions or different levels of data aggregation.

**D2d. Space-Time Scan Statistics for Early Detection of Disease Outbreaks**

All three proposed scan statistics extend readily to space-time tessellations. The space-time version of the elliptical scan statistic will employ cylindrical extensions of spatial ellipses rather than space-time ellipsoids, which would cause considerable analytical and computational complexity. The tree-structured and annealing scan statistics will examine arbitrarily shaped connected zones in space-time and for any user-specified definition of adjacency between space-time cells.

**D2e. Network Applications**

For certain problems, there is an underlying network structure on which we will want to perform the cluster detection and evaluation. For example, the New York City Health Department is monitoring the New York subway system and water distribution networks for bioterrorism attacks. In such a scenario, a circular scan statistic is not useful as two individuals close to each other in Euclidian distance may be very far from each other along the network. However, the tree-structured and annealing methods will be employed for the detection and evaluation of clusters on a predefined network. The essentially linear structure of these networks, compared with tessellation-derived networks, is expected to have a major impact on the form of the null distributions and their parametric approximations.
D3. Method Evaluation and Comparison

D3a. Optimal versus Semi-Optimal Clusters
By an “optimal” cluster is meant the zone that maximizes the profile likelihood across the full zonal parameter space $\Omega$, i.e., the true MLE. This true MLE will be computed by exhaustive search for a selection of small tessellations and the result compared with that produced by the three proposed scan statistics. However, it is not clear that these performance assessments extrapolate well to more realistic-sized tessellations where it is impossible to do an exhaustive search. Here, we will use the annealing-estimated cluster as the best currently available approximation to optimality to assess performance of the other two methods.

D3b. Power Evaluations
For power evaluation, we will complement the existing benchmark data sets described in section C6 with additional data sets based on alternative hypotheses with irregular shaped clusters such as along the Hudson River Valley.

D3c. Accuracy of Cluster Estimation
A central issue here is how to measure accuracy of estimation and several measures will be employed for this purpose. Simplest is the overlap probability, i.e., the probability that the estimated cluster has a nonempty intersection with the actual cluster. This is similar to power except that a power calculation/simulation records any declaration of a significant cluster as a “success” even when the cluster is not correctly located by the procedure. Going beyond mere overlap, we would like to measure the “distance” between actual and estimated cluster. For this, we will initially use the set-theoretic difference, in terms of either geographic area or population size. The level of aggregation and tessellation geometry should have major impacts upon this measure of accuracy.

D3d. Computation Time
Generally, we expect the tree-structured approach to be the fastest of the three methods and the annealing approach to be the slowest. Quantitative differences in timing among the methods as well as memory and data-input requirements will be documented for user-guidance purposes. An important consideration regarding timing is whether both high-rate and low-rate clusters are desired. The circular and elliptic scan statistics can identify both types of clusters in a single scan while the tree-structured and annealing approaches require two separate runs with a doubling of computer time.

D4. Data Sets and Applications

D4a. Breast Cancer Mortality in Northeastern United States
Compared to the United States as a whole, the Northeastern states have higher breast cancer mortality. In this study, we wanted to determine whether there were any statistically significant local clusters of excess breast cancer mortality within this already high area, or whether it was uniformly throughout. Using the spatial scan statistic, four different analyses were conducted, adjusting for age, age plus race, age plus urbanicity and age plus parity. The results were similar. For the age-adjusted analysis, there was one statistically significant cluster in the NYC–Philadelphia metropolitan area ($RR=1.07$, $p=0.0001$).

For additional information, see Kulldorff et al (1997).
D4b. Prostate Cancer Data in Maryland
Data from the Maryland Cancer Registry are available to the investigators, covering more than 24,000 incident cases of prostate cancer reported in Maryland during the years 1992–1997. These case records contain information on race, age, diagnosis, and disease stage and grade, and have been geo-coded to the coordinates (home addresses) of the cases as part of research on geographical analysis of prostate cancer in Maryland. Using SaTScan software, significant geographic variation in incidence, stage at diagnosis, and tumor grade have been detected for all cases, and for subgroups by race and age group. These data are well suited for testing new cluster detection methods, and for comparing cancer patterns detected by existing and new methods.

For additional information, see Klassen et al (2001) and Curriero et al (2001).

D4c. New York City Subway System Syndromic Surveillance Data
The New York City Department of Health (DOH) and Metropolitan Transportation Authority (MTA) began monitoring subway worker absenteeism in October 2001 as one of several surveillance systems for the early detection of disease outbreaks. Each day the MTA transmits an electronic line list of workers absent the previous day, including work location and reason for absence. DOH epidemiologists currently monitor temporal trends in absences in key syndrome categories (e.g., fever-flu or gastrointestinal illness). Analytic techniques are needed for detecting clustering within the subway network. These techniques must take into account the proximity of workers along the same subway line and the intersection of subway lines throughout the system. Data are currently stored at the DOH, and will be shared for the purposes of this research upon approval of a data-sharing protocol by the DOH, MTA and the principal investigator.

For additional information, Richard Heffernan of the New York City Department of Health.

D4d. Additional Data Sets
We will also, as needs arise for both testing and illustrative purposes, use other data sets selected from what we have worked with in the past, including brain cancer and prostate cancer mortality in the United States (Jemal et al 2002; Fang et al 2002), thyroid and brain cancer incidence in New Mexico (Kulldorff 2001; Kulldorff et al 1998), breast cancer incidence, staging and treatment in Connecticut (Gregorio et al 2002; Gregorio et al 2001) and Massachusetts (Sheehan et al 2000).

D5. Work Plan
D5a. Scope of the Project
With eight comprehensive live data bases for data testing and applications, five specific aims, three different spatial scan statistic techniques, and two types of data sets and models, the potential extent of this project is quite comprehensive with 30 sub-projects. As the work progresses, we will continuously use the results already obtained together with our scientific judgment to determine the specific choices and priorities for subsequent work, making sure that all of the 30 sub-projects will be thoroughly covered.

This comprehensive and intensive research proposal for geographic cancer surveillance and elevated rate area identification and interpretation is a multi-disciplinary, multi-project, multi-
individual and multi-site research undertaking. It is both comprehensive and intensive also because of its relevance to biological and chemical threats to societal networks and systems, such as NYC subway system. The proposed research requires a responsive well-integrated multi-site work plan. We will do the following for each year, internal to the project group:

1. **Comprehensive Integrative Workshop**: five-day hands on Summer workshop at Penn State.
2. **Elliptic and Simulated Annealing Spatial Scan Statistics and Northeast U.S. Breast Cancer Mortality Data Workshop**: three-day hands on Fall workshop at U Connecticut.
3. **Tree-Structured Spatial Scan Statistics for Space, Space-Time, and Networks in Theory and Practice**: four-day hands on Winter workshop at Penn State.

We will do the following for multi-project multi-investigator integration:

1. The project group will have in-house weekly MWF miniseminars and two-month-interspersed hands on thematic workshops, one at each performance site during a year. These regularized activities will strengthen the horizontal and vertical integration among the people and the projects proposed.
2. The project group will have the benefit of intimate internal review, comments, and advice on each manuscript in preparation for submission for publication and/or display of the material on the website.
3. The project group will give a short course in the third summer on geographic disease surveillance and hotspot detection and delineation using spatial scan statistics. This will help consolidate and strengthen both the project material and the project group further. The course will be planned in conjunction with a professional meeting, after the project’s internal summer-scheduled comprehensive workshop.
4. The project group will open up the last comprehensive summer workshop to the interested research community. This will help fine-tune and finalize the methods, tools, software, and the project materials to be prepared for final reports and final publications during the final fourth year of the project.

**D5b. Division of Labor**

There will be close interaction and collaboration between the project participants. While there are 30 separate sub-projects, these are very closely connected. For practical and logistic reasons, the project participants will take the lead on different sub-projects, as indicated in bold face in the following Table. All the participants will at the same time be actively involved with several other subprojects as listed in regular type.
Specific Aims | Poisson Model | Bernoulli Model
---|---|---

GP: Ganapati Patil | LP: Linda Pickle
CT: Charles Taillie | AK: Ann Klassen
MK: Martin Kulldorff | RH: Richard Heffernan
LD: Luiz Duczmal | PS1/PS2: Penn State Ph.D. Student #1/2
RA: Renato Assuncao | CS: Connecticut Ph.D. Student

*D5c. Synergy of Project Participants*
The project participants have worked together so far in two separate groups with Drs. Kulldorff and Patil in projects involving geospatial public health and geospatial ecosystem health. The promise of cross-disciplinary cross-fertilization for geographical disease surveillance has brought the two groups together to jointly develop the SaTScan methodology and technology for irregularly shaped clusters, and early warning system.

The project participants will use all of the available means of communication as seen productive and economical. The graduate students will be mobile as per need to keep the theory-data interactions moving throughout the project period. The project participants will also meet periodically at professional meetings and contribute to a broader interaction and synergism. The five performance sites will carry linked project websites for 24 hr contact, communication, interaction, and collaboration with central website at Penn State.

*D5d. Time Line*
Work on the 30 sub-projects will be done in a manner that is mutually synergistic. This is because, as much as the theoretical parts will inform and influence the practical ones, so will the practical and simulated benchmark data driven parts inform and influence the theoretical...
developments. While it is impossible to construct a simple work schedule diagram, the following outline may be suggestive of the work plan.

<table>
<thead>
<tr>
<th>Specific Aims</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Development of Novel Spatial Scan Statistics</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2. Generalization to Space-Time Versions</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3. Network Considerations and Variations</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>4. Comparisons with Simulated Benchmark Data</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5. Testing and Validation with Real Data</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

**D5e. Deliverables**

This project will result in several scientific publications that will be submitted to journals in the fields of theoretical, applied, and computational statistics, biometrics and biostatistics, cancer epidemiology, and public health. We expect to submit at least 6 papers per year. The software routines will be made available on the web for use by other scientists.

Special thematic issues of relevant journals will be explored on spatial scan statistics in theory and practice. The project group plans to prepare a monograph on methods and tools, and a casebook on data sets and applications soon after the successful completion of the project. Publishers’ interest is already on the horizon.

**E. Human Subjects**

This project involves secondary statistical analyses of existing data that are either publicly available (Northeast breast cancer mortality) or for which IRB approval has been obtained for the original data collection and analyses. Additional IRB approval will be sought for the new analyses to be conducted, and we will assure that the confidentiality of information about individuals is preserved. No new data will be collected.

The methods studied in this project can be used for data sets containing individuals of any gender, ethnicity and age group. The prostate cancer data include all ethnicities and all ages for which there was at least one case. The breast cancer data include all ethnicities and all ages but no men. Men are excluded since breast cancer in men is considered a different disease from that
in women, with different staging and treatment criteria. The subway worker data include all ethnicities and both genders.

**F. Vertebrate Animals**

None.

**G. Literature Cited**


Biomedware (2001). *Software for the Environmental and Health Sciences*. Biomedware, Ann Arbor, MI.


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


