

Upper level set scan statistic for detecting arbitrarily shaped hotspots

G. P. PATIL and C. TAILLIE

Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Penn State University, University Park, PA 16802

Received June 2002; Revised March 2003

A declared need is around for geoinformatic surveillance statistical science and software infrastructure for spatial and spatiotemporal hotspot detection. Hotspot means something unusual, anomaly, aberration, outbreak, elevated cluster, critical resource area, etc. The declared need may be for monitoring, etiology, management, or early warning. The responsible factors may be natural, accidental, or intentional.

This proof-of-concept paper suggests methods and tools for hotspot detection across geographic regions and across networks. The investigation proposes development of statistical methods and tools that have immediate potential for use in critical societal areas, such as public health and disease surveillance, ecosystem health, water resources and water services, transportation networks, persistent poverty typologies and trajectories, environmental justice, biosurveillance and biosecurity, among others.

We introduce, for multidisciplinary use, an innovation of the health-area-popular circle-based spatial and spatiotemporal scan statistic. Our innovation employs the notion of an upper level set, and is accordingly called the upper level set scan statistic, pointing to a sophisticated analytical and computational system as the next generation of the present day popular SaTScan.

Success of surveillance rests on potential elevated cluster detection capability. But the clusters can be of any shape, and cannot be captured only by circles. This is likely to give more of false alarms and more of false sense of security. What we need is capability to detect arbitrarily shaped clusters. The proposed upper level set scan statistic innovation is expected to fill this need

Keywords: confidence set of hotspots, early warning, geosurveillance statistics, hotspot detection, hotspot rating, nested upper level set scan statistic, typology of space-time hotspots

1352-8505 © 2004  Kluwer Academic Publishers

1. Introduction

Three central problems arise in geographical surveillance for a spatially distributed response variable. These are (i) identification of areas having exceptionally high (or low) response, (ii) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant, and (iii) assessment of explanatory factors that may account for the elevated response. Although a wide variety of methods have been proposed for modeling and analyzing spatial data (Cressie, 1991), the spatial scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) has quickly become a

1352-8505 © 2004  Kluwer Academic Publishers

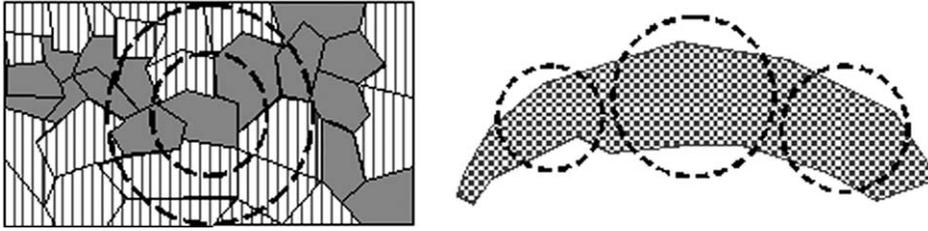


Figure 1. Limitations of circular scanning windows. (Left) An irregularly shaped cluster—perhaps a cholera outbreak along a winding river floodplain. Small circles miss much of the outbreak and large circles include many unwanted cells. (Right) Circular windows may report a single irregularly shaped cluster as a series of small clusters.

popular method for detection and evaluation of disease clusters, and is now widely used by many health departments, government scientists, and academic researchers. Kulldorff *et al.* (1998) have developed the SaTScan software system, which is available on the web without charge. A commercial software system (Biomedware, 2001) is also available. Two books (Glaz and Balakrishnan, 1999; Glaz *et al.*, 2001) cover the scan statistic, although their emphasis is on the one-dimensional version. When applied in space-time, the scan statistic approach can provide early warning of disease outbreaks and can monitor the spatial spread of an outbreak (Kulldorff, 2001; Mostashari *et al.*, 2002; Waller, 2002). With suitable modifications, the scan statistic approach can be used for critical area analysis in fields other than the health sciences. We describe some promising developments for generalizing the spatial scan statistic to make it applicable to hotspot-related issues encountered by environmental scientists.

Basic ingredients of the scan statistic are the geometry of the area being scanned, the probability distribution generating responses under the null-hypothesis of chance variation, and the shapes and sizes of the scanning window. Depending on the application, different response distributions are chosen and the test statistic is evaluated through Monte Carlo simulation (Dwass, 1957). Currently available spatial scan statistic software suffers from several limitations:

- First, circles have been used for the scanning window, resulting in low power for detection of irregularly shaped clusters (Fig. 1). Alternatively, an irregularly shaped cluster may be reported as a series of circular clusters. Kulldorff *et al.* (2002) explore the potential of elliptical scanning windows.
- Second, the response variable has been defined on the cells of a tessellated geographic region, preventing application to responses defined on a network (stream network, highway system, water distribution network, etc.).
- Finally, reflecting the epidemiological origins of the spatial scan statistic, response distributions have been taken as discrete (specifically, binomial or Poisson).

We suggest ways of overcoming all these limitations.

2. Background theory of scan statistics

The spatial scan statistic deals with the following situation. A region R of Euclidian space is tessellated or subdivided into cells (which will be denoted by the symbol a). Data are

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. 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:

- *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 $0 < p_a < 1$.
- *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, N_a people reside in cell a and each has a certain disease independently with probability p_a . The cell count Y_a is the number of diseased people in the cell. For the Poisson, A_a is the size (perhaps area or some adjusted population size) of the cell a , and Y_a is a realization of a Poisson process of intensity λ_a across the cell. In each scenario, the responses Y_a are independent; it is assumed that spatial variability can be accounted for by cell-to-cell variation in the model parameters.

The spatial scan statistic seeks to identify “hotspots” or “clusters” of cells that have an elevated response compared with the rest of the region. Elevated response means large values for the rates (or intensities),

$$G_a = \frac{Y_a}{A_a},$$

instead of for the raw counts Y_a . Cell counts are thus adjusted for cell sizes before comparing cell responses. The scan statistic easily accommodates other adjustments, such as for age or for gender.

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 geographically connected subset of the region R (Fig. 2). Such collections of cells will be referred to as zones and the set of all zones is denoted by Ω . Thus, a zone $Z \in \Omega$ is a 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 region.

The notion of a hotspot is inherently vague and lacks any *a priori* definition. There is no “true” hotspot in the statistical sense of a true parameter value. A hotspot is instead

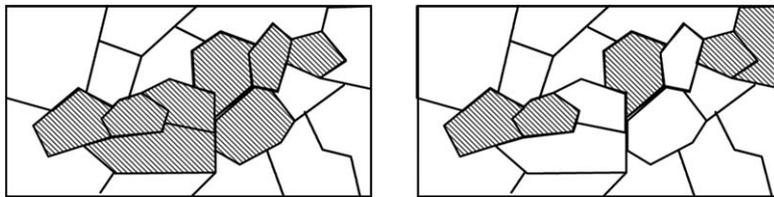


Figure 2. A tessellated region. The collection of shaded cells in the left-hand diagram is connected and, therefore, constitutes a zone in Ω . The collection on the right is not connected.

defined by its estimate—provided the estimate is statistically significant. To this end, the scan statistic adopts a hypothesis testing model in which the hotspot occurs as an unknown zonal parameter in the statement of the alternative hypothesis. The following is a statement of the null and alternative hypotheses in the binomial setting:

H_0 : p_a is the same for all cells in region R , i.e., there is no hotspot.

H_1 : There is a non-empty zone Z (connected union of cells) and parameter values $0 < p_0, p_1 < 1$ such that

$$p_a = \begin{cases} p_1 & \text{for all cells } a \text{ in } Z, \\ p_0 & \text{for all cells } a \text{ in } R - Z, \end{cases} \quad \text{and } p_1 > p_0.$$

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:

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

The null model, H_0 , is the limit of H_1 as $p_1 \rightarrow p_0$; however, the parameter Z is not identifiable in the limit. If one is searching for regions of low response, the condition $p_1 > p_0$ in the alternative hypothesis is changed to $p_1 < p_0$.

For given Z , the likelihood estimates of p_0 and p_1 can be written down explicitly which determines 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 Ω of all possible zones. In fact, Ω 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 across Ω_0 . This works well if Ω_0 contains the MLE for the full Ω or at least a close approximation to that MLE. Parameter space reduction is roughly analogous to doing a grid search in conventional optimization problems.
- (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. These procedures are computationally intensive enough that they can be difficult to replicate many times particularly when a simulation study is needed to determine null distributions. For this reason, stochastic optimization methods will not be discussed further in this paper; however, Duczmal and Assunção (2002) have applied simulated annealing to do global optimization for the scan statistic.

The traditional spatial scan statistic uses expanding circles to determine a reduced list Ω_0 of candidate zones Z . By their very construction, these candidate zones tend to be compact in shape and may do a poor job of approximating actual clusters. The circular scan statistic

has a reduced parameter space that is determined entirely by the geometry of the tessellation and does not involve the data in any way. The scan statistic that we propose takes an adaptive point of view in which Ω_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 $\Omega_0 = \Omega_{\text{ULS}}$ consists of all connected components of all upper level sets (ULS) of this surface. The cardinality of Ω_{ULS} does not exceed the number of cells in the tessellation. Furthermore, Ω_{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 Ω_{ULS} is data-dependent, this reduced parameter space must be recomputed for each replicate data set when simulating null distributions.

Although the traditional spatial scan statistic is applicable only to tessellated data, the ULS approach has an abstract graph (i.e., vertices and edges) as its starting point. Accordingly, this approach 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 vertices to be the cells of the tessellation. Two vertices are joined by an edge if the corresponding cells are adjacent in the tessellation. There is complete flexibility regarding the definition of adjacency. For example, one may declare two cells as adjacent (i) if their boundaries have at least one point in common, or (ii) if their common boundary has positive length, or (iii) in the case of a drainage network, if the flow is from one cell to the next. The user is free to adopt whatever definition of adjacency is most appropriate to the problem at hand.

3. ULS scan statistic

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

$$G_a = \frac{Y_a}{A_a}.$$

These rates determine a function $a \rightarrow G_a$ defined over the cells in the tessellation (more generally, the vertices of the abstract graph). This function has only finitely many values (levels) and each level g determines a ULS

$$U_g = \{a : G_a \geq g\}.$$

Since upper level sets do not have to be geographically connected (Fig. 3), we take the reduced list of candidate zones, Ω_{ULS} , to consist of all connected components of all possible upper level sets.

The zones in Ω_{ULS} are certainly plausible as potential hotspots since they are portions of upper level sets of the response rate. The number of zones is small enough for practical maximum likelihood search—in fact, the size of Ω_{ULS} does not exceed the number of vertices in the abstract graph (e.g., the number of cells in the tessellation). Finally, a tree structure can be defined on the reduced parameter space Ω_{ULS} . The nodes of the tree are the members of Ω_{ULS} , i.e., the candidate zones. Two nodes $Z, Z' \in \Omega_{\text{ULS}}$ are joined by an edge if

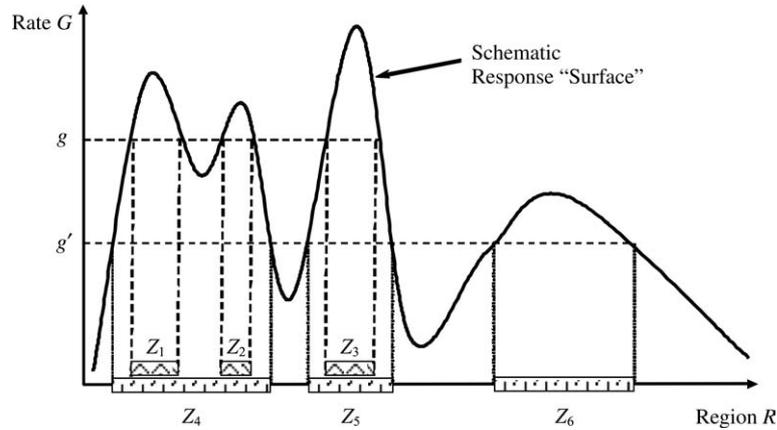


Figure 3. Schematic response surface with two response levels, g and g' . The upper level set determined by g has three connected components, Z_1 , Z_2 and Z_3 ; that determined by g' has Z_4 , Z_5 and Z_6 as its connected components. The diagram also illustrates the three ways in which connectivity can change as the level drops from g to g' : (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.

- i. Z is a proper subset of Z' , written as $Z \subsetneq Z'$.
- ii. There is no node $W \in \Omega_{\text{ULS}}$ such that $Z \subsetneq W \subsetneq Z'$.

This tree is called the ULS-tree; its nodes are the zones $Z \in \Omega_{\text{ULS}}$ and are therefore collections of vertices from the abstract graph. Leaf nodes are (typically) singleton vertices at which the response rate is a local maximum; we say “typically” because the response rate at a local maximum could be constant across several adjacent vertices. The root node consists of all vertices in the abstract graph (which we assume to be connected, for otherwise the ULS tree would be a forest instead of a tree). Fig. 4 shows the tree structure for the surface displayed in Fig. 3.

A consequence of adaptivity of the ULS approach is that Ω_{ULS} must be recalculated for each replicate in a simulation study. Efficient algorithms are needed for this calculation. 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 \mathbf{A} whose rows and columns are labeled with the cells of the tessellation. Entry \mathbf{A}_{ab} equals 1 if cells a and b are the same or are adjacent in the tessellation. Otherwise, \mathbf{A}_{ab} vanishes. The cells (row and column labels) are arranged in order of decreasing intensity $G_a = Y_a/A_a$ so that the adjacency matrix for any upper level set is a square submatrix in the northwest corner of the full adjacency matrix \mathbf{A} . This reordering of the rows and columns of \mathbf{A} is the only data dependent part of the algorithm. As the level drops cells are added one after another and one has to keep track of how the connectivity changes with each addition of a cell. As shown in Fig. 3,

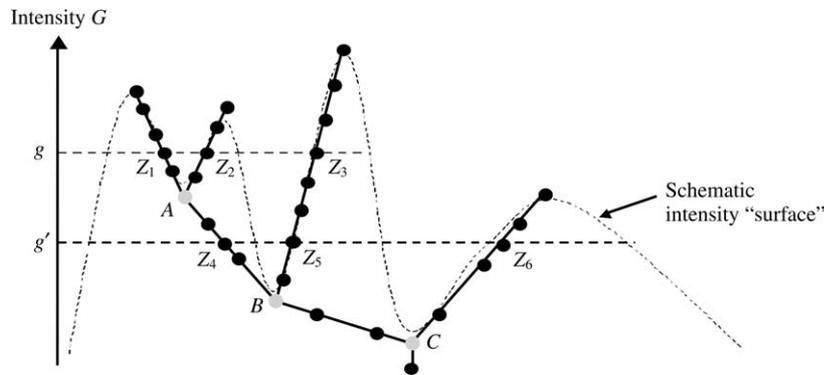


Figure 4. ULS connectivity tree for the schematic surface displayed in Fig. 3. The four leaf nodes correspond to surface peaks. The root node represents the entire region. Junction nodes (A, B and C) occur when two (or more) connected components coalesce into a single connected component.

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 $G_a = Y_a/A_a$. Computer efficiency can be further improved since it is not necessary to compute the portion of the tree below a specified level below which cells are background and not plausible locations for a hotspot.

4. Continuous response distributions

Our strategy for handling continuous responses is to model the mean and variance of each response distribution in term of the size variable A_a ; modeling is guided by the principle that the mean response should be proportional to A_a and the relative variability should decrease with A_a . Just as with the Poisson and binomial models, we take the Y_a to be independent. The approach is best illustrated for the gamma family of distributions.

4.1 Gamma distribution

We parameterize the gamma distribution by (k, β) where k is the index parameter and β is the scale parameter. Thus, if Y is a gamma-distributed variate,

$$E[Y] = k\beta \quad \text{and} \quad \text{Var}[Y] = k\beta^2.$$

Both k and β can vary from cell to cell but additivity with respect to the index parameter suggests that we take k proportional to the size variable:

$$k_a = \frac{A_a}{c},$$

where c is an unknown parameter but whose value is the same for all a . This gives the following mean and squared coefficient of variation:

$$E[Y_a] = \frac{\beta_a A_a}{c} \quad \text{and} \quad \text{CV}^2[Y_a] = \frac{c}{A_a}.$$

The hotspot hypothesis testing model is analogous to that of the binomial described previously:

H_0 : β_a is the same for all a , i.e., there is no hotspot.

H_1 : There is a non-empty zone Z and parameter values $0 < \beta_0, \beta_1$ such that

$$\beta_a = \begin{cases} \beta_1 & \text{for all objects } a \text{ in } Z, \\ \beta_0 & \text{for all objects } a \text{ outside of } Z, \end{cases} \quad \text{and} \quad \beta_1 > \beta_0.$$

The full model has four unknown parameters Z, c, β_0, β_1 that need to be estimated. The profile likelihood function for Z is obtained by fixing an arbitrary candidate zone Z and maximizing the likelihood with respect to the other three parameters. The latter optimization problem reduces to the solution of three likelihood equations. Two of these equations can be solved for β_0 and β_1 in terms of c :

$$\beta_0 = c \frac{\sum^{(0)} Y_a}{\sum^{(0)} A_a} \quad \text{and} \quad \beta_1 = c \frac{\sum^{(1)} Y_a}{\sum^{(1)} A_a}, \quad (1)$$

where $\sum^{(0)}$ and $\sum^{(1)}$ indicate summation over objects outside Z and inside Z , respectively. Equations (1) are used to eliminate β_0 and β_1 from the remaining likelihood equation, giving a single equation to be solved for the final parameter c . This equation, which cannot be solved in closed form, is

$$\begin{aligned} \sum_a A_a \left[\log\left(\frac{A_a}{c}\right) - \psi\left(\frac{A_a}{c}\right) \right] &= \left(\sum^{(0)} A_a \right) \log \frac{\sum^{(0)} Y_a}{\sum^{(0)} A_a} \\ &+ \left(\sum^{(1)} A_a \right) \log \frac{\sum^{(1)} Y_a}{\sum^{(1)} A_a} - \sum_a A_a \log \frac{Y_a}{A_a}, \end{aligned}$$

where $\psi(\cdot)$ is the digamma function. It is well known that the function

$$g(t) = \log(t) - \psi(t), \quad t \geq 0,$$

is strictly increasing with $g(0) = 0$ and $g(\infty) = \infty$. Accordingly, the LHS of Equation (2) is a strictly decreasing function of c which ranges from ∞ down to 0. On the other hand, the RHS of (2) is non-negative since arithmetic means are greater than geometric means. Thus, Equation (2) gives a unique MLE for c . The Newton–Raphson algorithm gives rapid

convergence. The likelihood estimate for the hotspot zone Z is obtained by maximizing the profile likelihood on the ULS tree, as before.

4.2 Lognormal and other continuous distributions

A similar approach is applicable to other two-parameter families of distributions on the positive real line. Specifically, for the lognormal distribution, we take

$$E[Y_a] = \frac{\beta_a A_a}{c} \quad \text{and} \quad \text{CV}^2[Y_a] = \left[\frac{c}{A_a} \right]^d,$$

where d is either user-specified (e.g., $d = 1$) or is an unknown parameter to be estimated. In terms of its conventional parameters (μ, σ^2) , the first two moments of the lognormal are

$$E[Y] = e^{\mu + \sigma^2/2} \quad \text{and} \quad \text{CV}^2[Y] = e^{\sigma^2} - 1,$$

which gives

$$e^{\mu_a} = \frac{A_a/c}{\sqrt{1 + (c/A_a)^d}} \beta_a \quad \text{and} \quad e^{\sigma_a^2} = 1 + \left(\frac{c}{A_a} \right)^d.$$

These equations explicitly specify the lognormal parameters (μ, σ^2) for each a in terms of the unknown parameters so that the likelihood can be written down explicitly (assuming independence).

4.3 Simulating the null distribution to obtain p -values

Conditional simulation is used to obtain the null distribution in the cases of the binomial and Poisson response distributions. One conditions on the sufficient statistic (under H_0) to eliminate the unknown parameters from the null model. The resulting parameter-free distributions are hypergeometric and multinomial, respectively, and are easily simulated. This is not the case for most continuous distributions. A glance at the H_0 -version of Equation (2) shows that one of the sufficient statistics is $\sum A_a \log(Y_a/A_a)$ and the conditional distribution is not anything familiar. Accordingly, simulation might be done by replacing unknown parameters with their maximum likelihood estimates under H_0 .

5. 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.

5.1 Hotspot membership rating

Extending this idea, zonal estimation uncertainty can be visually depicted by inner and outer envelopes, where the outer envelope consists of all cells belonging to at least one zone in the confidence set. Cells in the inner envelope belong to all (or to a sufficiently large percentage) of the zones in the confidence set. In other words, the outer envelope is the union of all zones in the confidence set while the inner envelope is their intersection (Fig. 5).

The hotspot confidence set also lets us assign a numerical rating to each cell for inclusion in the hotspot. The rating is the percentage of zones in the confidence set that includes the cell under consideration. The inner envelope consists of cells receiving a 100% rating while the outer envelope contains the cells with a nonzero rating. A map of these ratings, with superimposed MLE, provides a visual display of uncertainty in hotspot delineation.

5.2 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 α (Bickel and Doksum, 1977, p. 179; Lehmann, 1986, Section 3.5, Theorem 4). The confidence level is $c = 1 - \alpha$. Alternatively, the confidence set contains all null hypotheses for which the p -value exceeds $1 - c$. In the present setting, the parameter space is Ω_{ULS} : the set of all zones which are connected components of upper level sets of the rate function. The confidence set will be a subset of Ω_{ULS} and a particular zone $Z_0 \in \Omega_{\text{ULS}}$ is in the confidence set if we cannot reject the following null hypothesis (formulated for the binomial distribution for definiteness):

\tilde{H}_0 : There are binomial parameters $p_1 \geq p_0$ such that

$$p_a = \begin{cases} p_1, & \text{for all cells } a \text{ in } Z_0; \\ p_0, & \text{for all cells } a \text{ outside } Z_0. \end{cases}$$

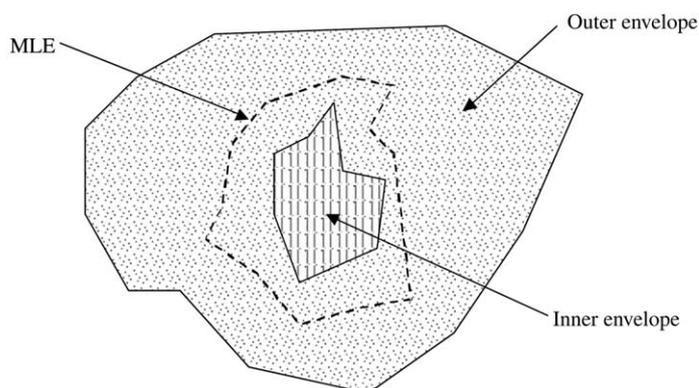


Figure 5. Estimation uncertainty in hotspot delineation. Cells in the inner envelope belong to all plausible estimates (at specified confidence level); cells in the outer envelope belong to at least one plausible estimate. The MLE is nested between the two envelopes.

This hypothesis is to be tested against the general alternative in which the hotspot zone Z is allowed to vary freely over Ω_{ULS} . Schematically, then, we are testing $\tilde{H}_0 : Z = Z_0$ versus $\tilde{H}_1 : Z \neq Z_0$ and the confidence set consists of all zones Z_0 for which \tilde{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 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 \tilde{H}_0 involves, as nuisance parameters, the rates p_0 and p_1 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 ULS tree structure on our reduced parameter space Ω_{ULS} . The nodes of Ω_{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 (Fig. 6). As shown in Fig. 6, 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

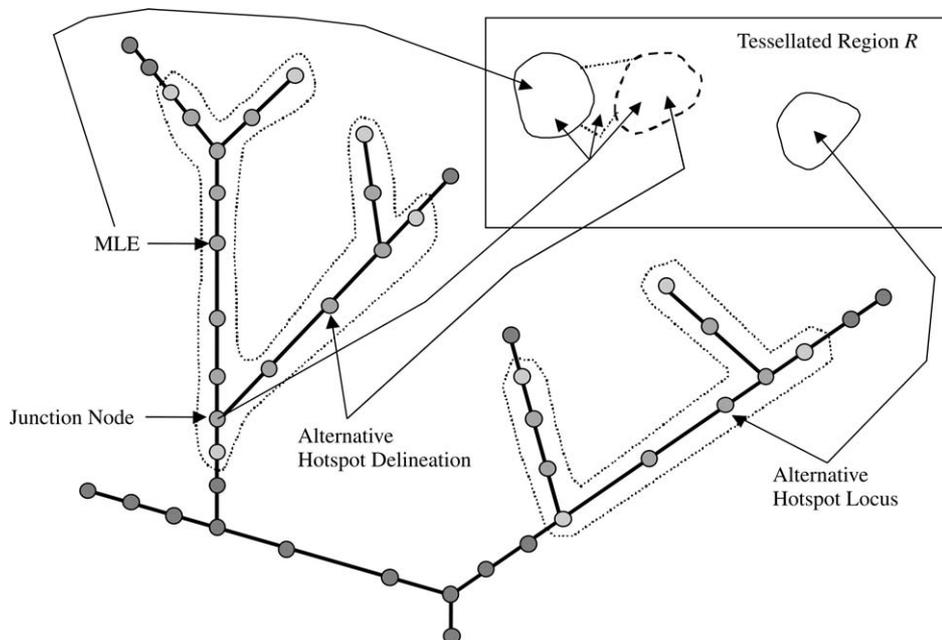


Figure 6. A confidence set of hotspots on the ULS tree. The different connected components correspond to different hotspot loci while the nodes within a connected component correspond to different delineations of that hotspot—all at the appropriate confidence level.

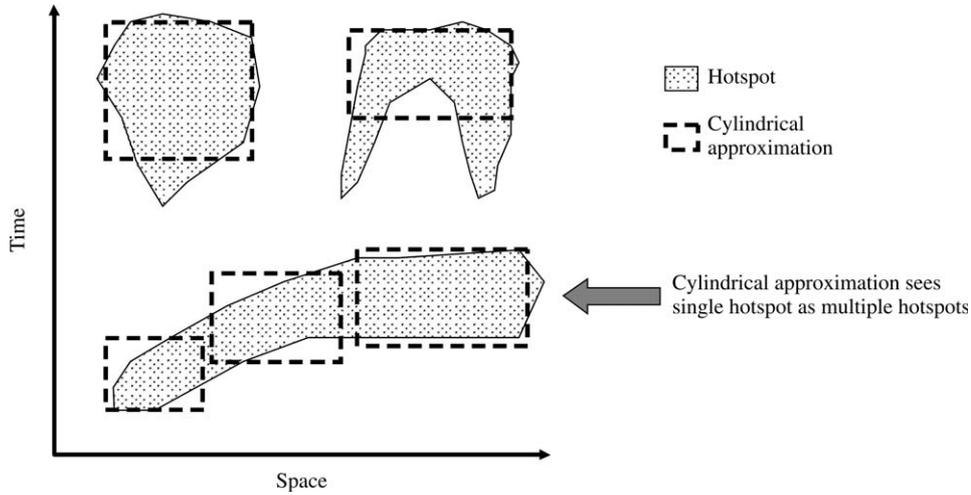


Figure 7. Temporal evolution of a spatial hotspot is represented by the shape of the hotspot in space-time. Cylinders may not adequately capture this shape.

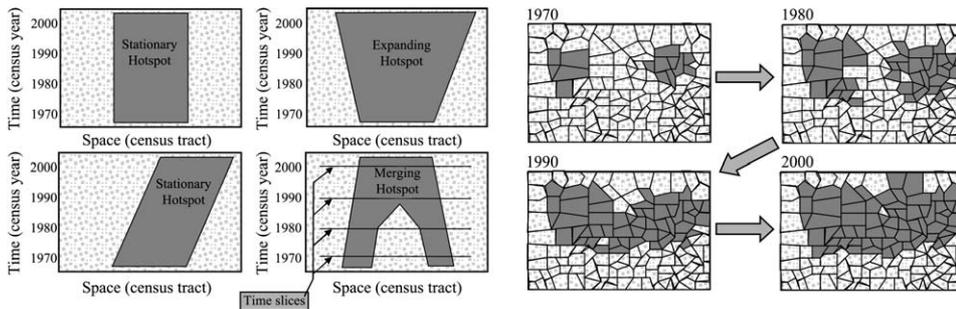


Figure 8. The four diagrams on the left depict different types of space-time hotspots. The spatial dimension is represented schematically on the horizontal axis while time is on the vertical axis. The diagrams on the right show the trajectory (sequence of time slices) of a merging hotspot.

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.

6. Filtering for explanatory variables

The scan statistic searches for regions of high response relative to a geo-referenced set of prior expected responses. Thus, a hotspot map depicts regions of extreme departure from expectation in the multiplicative sense, i.e., multiplicative residuals. The size values A_a , which are proportional to model expectations, are the link between the response variable and potential explanatory variables. In disease surveillance, the A_a are routinely adjusted for factors like age, gender, and population size before beginning the analysis (Bithell *et*

al., 1995; Kulldorff *et al.*, 1997; Rogerson, 2001; Waller, 2002; Walsh and Fenster, 1997; Walsh and DeChello, 2001). Such standard, agreed-upon, factors are often unavailable in other applications in which case the initial analysis may identify absolute hotspots by setting all A_a equal to unity. Locations of these highs can provide clues for identifying potential explanatory factors. Next, the size values are adjusted for these factors and the scan statistic is rerun with the adjusted sizes. Comparative configuration of new and old hotspots reveals the impact of these factors upon the response under study.

Several methods are available for adjusting the A_a . Suppose, first, that there is only one explanatory variable X . A nonparametric approach partitions the X -values into intervals and calculates the mean response for each interval. These calculations should utilize all available pertinent data. The adjusted size value for vertex a becomes

$$A'_a = \frac{m_a}{m} A_a,$$

where A_a is the old size value, m_a is the mean response for the interval containing vertex a , and m is an overall mean response. Regression of Y upon X can also be the basis for adjustment provided an appropriate functional relation is identified. Similar approaches work, in principle, for multiple factors. However, the “curse of dimensionality” often comes into play and data sparseness prevents calculation of dependable local means. Our approach, in such cases, is to cluster the data points in factor space. A mean response is then calculated for each cluster.

7. Typology of space-time hotspots

Scan statistic methods extend readily to the detection of hotspots in space-time. The space-time version of the circle-based scan statistic employs cylindrical extensions of spatial circles. But cylinders are often unable to adequately represent the temporal evolution of a hotspot (Fig. 7). The space-time generalization of the ULS scan statistic can detect arbitrarily shaped hotspots in space-time. This lets us classify space-time hotspots into various evolutionary types—a few of which appear on the left hand side of Fig. 8. The merging hotspot is particularly interesting because, while it comprises a connected zone in space-time, several of its time slices are spatially disconnected. The diagrams in Fig. 8 are motivated a study on “trajectories of persistent poverty in the US” being conducted by Amy Glasmeier of Penn State University. Census tract data for the 1970–2000 census years are used in the study.

8. Some additional issues

Nested ULS scan statistic. The hypothesis-testing model for the spatial scan statistic supposes that the response rate takes only two distinct values—an elevated value in the hotspot zone Z and a smaller value outside Z . This is a very crude approximation since the response rate actually varies gradually from location to location. A more realistic model might use nested zones $Z_1 \subseteq Z_2$ in which the response rate takes a very high value in Z_1 , a moderately high value in $Z_2 - Z_1$, and a low value outside Z_2 . The zones Z_1 and Z_2 are unknown model parameters and need to be estimated. Maximizing the likelihood function

across all nested pairs $Z_1 \subseteq Z_2$ would be computationally infeasible with the circles approach because of the large number of such pairs. Their number becomes much more practical when Z_1 and Z_2 are restricted to nodes on the ULS tree. In addition, the search can be limited to portions of the ULS tree in the vicinity of the single-zone MLE.

Parameterized null distributions. Under standard conditions, the log-likelihood ratio statistic is asymptotically chi-squared with appropriate degrees of freedom. Unfortunately, the scan statistic setting is highly non-standard since (i) the zonal parameter space Ω is finite and discrete and (ii) the zonal parameter is not identifiable under the null hypothesis (see Davies, 1977, in this connection). Nonetheless, it is natural to ask if 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 whether the circle-based or ULS-scan statistic is used. No general *a priori* rules relating parameter values to these conditions can be expected, so parameters for approximating null distributions will be estimated using simulated data. But, this should reduce substantially the number of replicates required and should also allow extrapolation to smaller p -values. Also, fitted null distributions and their parameters are of independent interest for characterizing and contrasting different geographical regions or different levels of data aggregation.

Acknowledgments

Prepared with partial support from U.S. EPA Star Grant for Atlantic Slope Consortium, Cooperative Agreement Number R-82868401, and the National Science Foundation Digital Government Program Award Number EIA-0307010. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agencies and no official endorsement should be inferred.

References

- Aarts, E. and Korst, J. (1989) *Simulated Annealing and Boltzmann Machines*, Wiley, Chichester.
- Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.
- Bithell, J.F., Dutton, S.J., Neary, N.M., and Vincent, T.J. (1995) Controlling for socioeconomic confounding using regression methods. *Community Health*, **49**, S15–S19.
- Cormen, T.H., Leieron, C.E., Rivest, R.L., and Stein, C. (2001) *Introduction to Algorithms* (second edition), MIT Press, Cambridge, Massachusetts.
- Cressie, N. (1991) *Statistics for Spatial Data*, Wiley, New York.
- Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–54.
- Duczmal, L. and Assunção, R.A. (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, in press.

- Dwass, M. (1957) Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181–7.
- Glaz, J. and Balakrishnan, N. (eds) (1999) *Scan Statistics and Applications*, Birkhauser, Boston.
- Glaz, J., Naus, J., and Wallenstein, S. (2001) *Scan Statistics*, Springer-Verlag, New York.
- Knjazew, D. (2002) *OmeGA: A Competent Genetic Algorithm for Solving Permutation and Scheduling Problems*, Kluwer Academic Publishers, Boston, Massachusetts.
- Knuth, D.E. (1973) *The Art of Computer Programming: Volume 1, Fundamental Algorithms*, (second edition), Addison-Wesley, Reading, Massachusetts.
- Kulldorff, M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–96.
- Kulldorff, M. (2001) Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, **164**, 61–72.
- 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**, 161–70.
- Kulldorff, M., Huang, L., and Pickle, L. (2004) An elliptic spatial scan statistic. *Manuscript, to be submitted*.
- Kulldorff, M. and Nagarwalla, N. (1995) Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14**, 799–810.
- Kulldorff, M., Rand, K., Gherman, G., Williams, G., and DeFrancesco, D. (1998) SaTScan version 2.1: Software for the spatial and space-time scan statistics. National Cancer Institute, Bethesda, MD.
- Lehmann, E.L. (1986) *Testing Statistical Hypotheses* (second edition), Wiley, New York.
- Mostashari, F., Kulldorff, M., and Miller, J. (2002) Dead bird clustering: An early warning system for West Nile virus activity. (Manuscript prepared for the New York City West Nile Virus Surveillance Working Group.) Under review.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) *Numerical Recipes in C* (second edition), Cambridge University Press, Cambridge.
- Rogerson, P.A. (2001) Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A*, **164**, 87–96.
- Waller, L. (2002) Methods for detecting disease clustering in time or space. In *Statistical Methods and Principles in Public Health Surveillance*, R. Brookmeyer and D. Stroup (eds), Oxford University Press, Oxford.
- Walsh, S.J. and DeChello, L.M. (2001) Geographical variation in mortality from systemic lupus erythematosus in the United States. *Lupus*, **10**, 637–46.
- Walsh, S.J. and Fenster, J.R. (1997) Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, 1981–1990. *The Journal of Rheumatology*, **24**, 2348–52.
- Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Springer, New York.

Biographical sketches

G. P. Patil is a Distinguished Professor of Mathematical Statistics and Director of the Center for Statistical Ecology and Environmental Statistics in the Department of Statistics at The Pennsylvania State University.

C. Taillie is a Senior Research Associate in the Center for Statistical Ecology and Environmental Statistics in the Department of Statistics at The Pennsylvania State University.