

Geographic and Network Surveillance via Scan Statistics for Critical Area Detection

G. P. Patil and C. Taillie

Abstract. Both statistical ecology and environmental statistics have numerous challenges and opportunities in the waiting for the twenty-first century, calling for increasing numbers of nontraditional statistical approaches. Both theoretical and applied ecology are using advancing data analytical and interpretational software and hardware to satisfy public policy and discovery research, variously incorporating geospatial information, site-specific data and remote sensing imagery. We discuss a declared need for geoinformatic surveillance for spatial critical area detection. We explore, for ecological and environmental use, an innovation of the circle-based spatial scan statistic popular in the health sciences.

Key words and phrases: Geoinformatic surveillance, hot-spot detection, Monte Carlo hypothesis testing, upper level set, upper level set scan statistic.

1. INTRODUCTION

Ecological and environmental studies are undergoing major changes in response to changing societal concerns coupled with remote sensing information and computer technology. Both theoretical and applied ecology are using more statistical thought processes and procedures with advancing software and hardware to satisfy public policy and research, variously incorporating geospatial information, sample survey data, intensive site-specific data and remote sensing image data. The issues are calling for increasing numbers of nontraditional statistical approaches (Patil, 1996). Both statistical ecology and environmental statistics have numerous challenges and opportunities in the waiting for the twenty-first century. While much progress has been made in the past, the future promises even more rapid developments as sophisticated computing

technology is utilized to apply newly developed statistical methods to increasingly detailed databases in both space and time in response to the demands of both policy and discovery. See, for example, Johnson and Patil (2004), Myers and Patil (2004), Myers and Patil (2002), Patil (2002), Patil et al. (2002), Patil et al. (2001) and Patil, Johnson, Myers and Taillie (2000).

In this article, we highlight landscape scales in statistical ecology, environmental statistics and geospatial risk assessment. There is a declared need for geoinformatic surveillance for geospatial hot-spot detection. Hot-spot means an anomaly, aberration, outbreak, elevated cluster, critical resource area and so on. The declared need may be for monitoring, etiology, management or early warning in critical societal areas, such as ecosystem health, water resources and water services, stream and transportation networks, persistent poverty typologies and trajectories, public health and disease surveillance, environmental justice, biosurveillance and biosecurity, among others. The responsible factors may be natural, accidental or intentional.

We discuss, for ecological and environmental use, an innovation of the circle-based spatial scan statistic (Kulldorff, 1997; Patil et al., 2002) popular in health science. Our innovation employs the notion of an upper-level-set based scan and is accordingly called the upper level set scan statistic, pointing to a sophisticated analytical and computational system as the

G. P. Patil is Distinguished Professor and Director, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802 (e-mail: gpp@stat.psu.edu). C. Taillie is Senior Research Associate, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802.

1 next generation of the present day SaTScan (Kulldorff,
 2 1997; Patil et al., 2002).

3
 4 **2. CRITICAL AREA DETECTION WITH THE**
 5 **SPATIAL SCAN STATISTIC**

6 Three central problems arise in geographical sur-
 7 veillance for a spatially distributed response variable.
 8 These are (i) identification of areas having excep-
 9 tionally high (or low) response, (ii) determination of
 10 whether the elevated response can be attributed to
 11 chance variation (false alarm) or is statistically sig-
 12 nificant and (iii) assessment of explanatory factors
 13 that may account for the elevated response. Although
 14 a wide variety of methods have been proposed for
 15 modeling and analyzing spatial data (Cressie, 1991),
 16 the spatial scan statistic (Kulldorff and Nagarwalla,
 17 1995; Kulldorff, 1997) has quickly become a pop-
 18 ular method for detection and evaluation of disease
 19 clusters and is now widely used by many health de-
 20 partments, government scientists and academic re-
 21 searchers (Kulldorff et al., 1998a; Kulldorff et al.,
 22 1998b; Kulldorff, 2001). With suitable modifications,
 23 the scan statistic approach can be used for critical area
 24 analysis in fields other than the health sciences. We de-
 25 scribe some promising developments for generalizing
 26 the spatial scan statistic to make it applicable to many
 27 issues in environmental science.

28 As in all geospatial surveillance, it is important to
 29 determine whether any variation observed may reason-
 30 ably be due to chance or not. This can be done using
 31 tests for spatial randomness, adjusting for the uneven
 32 geographical population density as well as for age and
 33 other known risk factors. One such test is the spatial
 34 scan statistic, which is used for the detection and eval-
 35 uation of local clusters or hot-spot areas. This method
 36 is now in common use by various governmental health
 37 agencies, including the National Institutes of Health,
 38 the Centers for Disease Control and Prevention and
 39 the state health departments in New York, Connecticut,
 40 Texas, Washington, Maryland, California and New Jer-
 41 sey. Other test statistics are more global in nature, eval-
 42 uating whether there is clustering in general throughout
 43 the map, without pinpointing the specific location of
 44 high or low incidence or mortality areas.

45 The spatial scan statistic has been implemented in
 46 two statistical software packages. One of these is the
 47 freely available SaTScan software (Kulldorff et al.,
 48 1998b) that was developed by and is distributed by the
 49 National Cancer Institute. The other is the ClusterSeer
 50 software (BioMedware, 2001), a commercial product.

52 **3. SCAN STATISTIC SUCCESS STORIES**

53 The circular spatial scan statistic and the accompa-
 54 nying SaTScan software are widely used by both gov-
 55 ernmental health departments and academic epidemi-
 56 ologists. Some of the past and present applications in-
 57 clude the following:

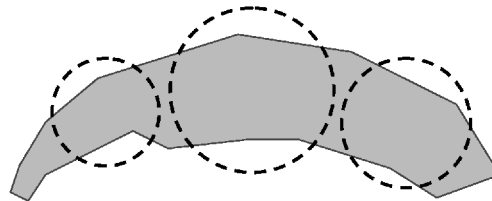
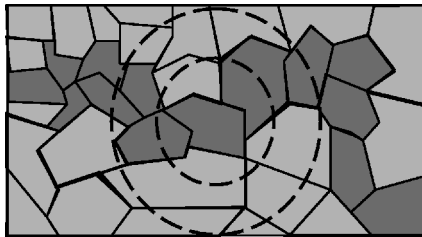
- 58 • *New York City Health Department*—Daily surveil-
 59 lance for the early detection of disease outbreaks.
 60 During the summer of 2001 it was successfully
 61 used for the early detection of dead bird clusters
 62 to quickly detect local West Nile virus epicenters.
 63 Cluster findings led to preventive measures such as
 64 targeted application of mosquito larvicide. During
 65 the spring of 2001 SaTScan was successfully used
 66 as the early detection tool in a simulated bioterror-
 67 ism exercise to train the New York City mayor, his
 68 staff and health department officials in emergency
 69 preparedness and conduct. Currently it is used for
 70 daily syndromic surveillance based on 911 emer-
 71 gency calls and hospital emergency admissions. For
 72 additional information, see Mostashari, Kulldorf and
 73 Miller (2002).
 74
- 75 • *Washington State Health Department*—Evaluation
 76 of a glioblastoma cluster alarm around Seattle–
 77 Tacoma International Airport. Earlier analyses had
 78 been inconclusive as results depended on geograph-
 79 ical boundaries chosen to define this cancer cluster,
 80 and there were also questions concerning preselec-
 81 tion bias of airport area when testing the difference
 82 in the incidence rate close to the airport versus fur-
 83 ther away from the airport. A SaTScan analysis for
 84 the county as a whole revealed a nonsignificant clus-
 85 ter around the airport, adding weight to other ev-
 86 idence that it was probably a chance occurrence.
 87 For additional information, see VanEenwyk et al.
 88 (1999).
- 89 • *National Creutzfeldt–Jakob Disease Surveillance*
 90 *Unit and the Leicester Health Authority, England*—
 91 A very small but statistically significant ($p = 0.004$)
 92 cluster with five cases of Creutzfeldt–Jakob disease
 93 was found in Charnwood, Leicestershire, England.
 94 A detailed local epidemiological investigation iden-
 95 tified specific and unusual butcher shop practices as
 96 the likely cause for this cluster. For additional infor-
 97 mation, see Bryant and Monk (2001), Cousens et al.
 98 (2001) and d’Aignaux et al. (2002).

99 **4. PROPERTIES OF THE SCAN STATISTIC**

100 The scan statistic is a statistical method with many
 101 potential applications, designed to detect a local excess
 102

1 of events and to test if such an excess can reasonably
 2 have occurred by chance. The scan statistic was first
 3 studied in detail by Naus (1965a, b), who looked at
 4 the problem in both one and two dimensions. Glaz,
 5 Naus and Wallenstein (2001) recently published a book
 6 summarizing the field, complementing an earlier edited
 7 volume (Glaz and Balakrishnan, 1999). In two or more
 8 dimensions, the events may be cases of leukemia, with
 9 an interest to see if there are geographical clusters of
 10 the disease; they may be antipersonnel mines, with
 11 an interest to detect large mine fields for removal;
 12 they could be Geiger counts, with an interest to detect
 13 large uranium deposits; they could be stars or galaxies;
 14 they could be breast calcifications showing up in
 15 a mammogram, possibly indicating a breast tumor;
 16 or they could be a particular type of archaeological
 17 pottery.

18 Three basic properties of the scan statistic are the
 19 geometry of the area being scanned, the probability
 20 distribution generating events under the null hypothesis
 21 and the shapes and sizes of the scanning window.
 22 Depending on the application, different models
 23 are chosen, and depending on the model, the test
 24 statistic is evaluated either through explicit mathematical
 25 derivations and approximations or through Monte
 26 Carlo sampling (Dwass, 1957). Due to inhomogeneous
 27 geographical population densities, there are no known
 28 asymptotic or approximate solutions for most disease
 29 surveillance problems, and Monte Carlo sampling is
 30 then used. Random data sets are generated under the
 31 known null hypothesis, and the value of the scan
 32 statistic is calculated for both the real data set and the
 33 simulated random data sets; if the former is among the
 34 5% highest, then the detected cluster is significant at
 35 the 0.05 level. While computer intensive, the Monte
 36 Carlo approach is quite feasible, and it is possible to
 37 analyze data sets with 10,000 + geographical locations
 38 and 100,000 cases or more.



49 FIG. 1. *Limitations of circular scanning windows: (left) an irregularly shaped cluster—perhaps a cholera outbreak along a winding river*
 50 *floodplain; small circles miss much of the outbreak and large circles include many unwanted cells; (right) circular windows may report a*
 51 *single irregularly shaped cluster as a series of small clusters.*

Multidimensional scan statistics have been studied 52
 for a long time. In terms of the region being scanned, 53
 Naus (1965b), Loader (1991), Alm (1997, 1998) and 54
 Anderson and Titterington (1997) all considered a two- 55
 dimensional rectangle. Alm (1998) also looked at a 56
 three-dimensional rectangular volume. Chen and Glaz 57
 (1996) studied a regular grid of discrete points within 58
 a rectangular area. Turnbull et al. (1990) used an 59
 irregular grid, where points may be anywhere within 60
 an arbitrarily shaped area. 61

Under the null hypothesis, Naus (1965b), Loader 62
 (1991) and Alm (1997, 1998) looked at a homogeneous 63
 Poisson process, Turnbull et al. (1990) considered an 64
 inhomogeneous Poisson process, and Anderson and 65
 Titterington (1997) considered both types. Chen and 66
 Glaz (1996) considered a Bernoulli model. As for the 67
 scanning window, Naus (1965b), Loader (1991), Chen 68
 and Glaz (1996), Alm (1997, 1998) and Anderson 69
 and Titterington (1997) all considered rectangles. Alm 70
 (1997, 1998) also looked at circles, triangles and other 71
 convex shapes. Turnbull et al. (1990) considered a cir- 72
 cular window centered at any of the grid points making 73
 up the data. The window is, in all cases, of fixed shape 74
 as well as of fixed size in terms of the expected number 75
 of events, with the exception of Loader (1991), who 76
 also considered a variable-size window. Based on the 77
 likelihood ratio test, Kulldorff (1997) presented a gen- 78
 eral mathematical model that includes all these cases, 79
 but even with the use of Monte Carlo sampling, it is 80
 not always computationally feasible to evaluate all pos- 81
 sible window locations, sizes and shapes. While we no 82
 longer have to worry about the very difficult mathemat- 83
 ics entailed in finding approximate or asymptotic solu- 84
 tions, we must now worry about efficient algorithms 85
 for evaluating a very large number of windows. 86

Currently available spatial scan statistic software has 87
 several limitations. First, circles have been used for the 88
 scanning window, resulting in low power for detection 89
 of irregularly shaped clusters (Figure 1). Alternatively, 90

91
 92
 93
 94
 95
 96
 97
 98
 99
 100
 101
 102

1 an irregularly shaped cluster may be reported as a series of circular shaped clusters. 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 some ways of addressing these limitations.

12 **5. BASIC THEORY OF THE SCAN STATISTIC**

13 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 nonnegative counts Y_a on cells 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:

- 23 • *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 $0 < p_a < 1$.
- 26 • *Poisson*— A_a is a positive real number and $Y_a \sim \text{Poisson}(\lambda, A_a)$, where $\lambda_a > 0$ is an unknown parameter attached to cell a .

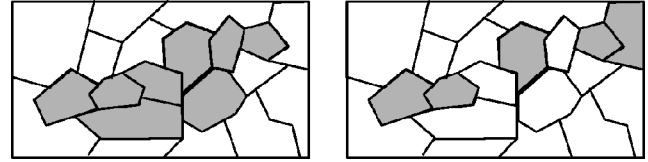
29 Each distributional model has a simple interpretation. For the binomial, N_a people reside in cell a and each has a certain disease independently with probability p_a . The cell count Y_a is the number of diseased people in the cell. For the Poisson, A_a is the size (perhaps area) of the cell a , 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.

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

$$41 \quad G_a = Y_a/A_a,$$

46 instead of for the raw counts Y_a . In other words, 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.

51 A collection of cells from the tessellation should satisfy several geometrical properties before it could be



52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102

FIG. 2. A tessellated region: the collection of shaded cells in the left diagram is connected and, therefore, constitutes a zone in Ω ; the collection on the right is not connected.

considered as a candidate for a hot-spot cluster. First, the union of the cells should comprise a geographically connected subset of the region R (Figure 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 hot spots to zones that do not comprise more than, say, 50% of the region.

The notion of a hot spot is inherently vague and lacks any a priori definition. There is no “true” hot spot in the statistical sense of a true parameter value. A hot spot is instead defined by its estimate—provided the estimate is statistically significant. The scan statistic adopts a hypothesis testing model in which the hot spot 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 , that is, there is no hot spot.

H_1 : There is a nonempty 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:

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

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 p_0 of p_1 and can be written 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 hot-spot 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. See Duczmal and Assuncao (2003).

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 piecewise 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 hot-spot 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.

6. UPPER LEVEL SET SCAN STATISTIC

The upper level set 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

$$G_a = Y_a/A_a.$$

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

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

Since upper level sets do not have to be geographically connected, the reduced list of candidate zones, Ω_{ULS} , consists of all connected components of all possible upper level sets.

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 (see Cormen, Leieron, Rivest and Stein, 2001, Section 22.3, for depth first search; Knuth, 1973, page 353; or Press, Teukolsky, Vetterling and Flannery, 1992, Section 8.6, for transitive closure).

6.1 Continuous Response Distributions

Our strategy for handling continuous responses is to model the mean and variance of each response distribution in terms 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.

Gamma distribution. We parameterize the gamma distribution by (k, β) , where k is the index parameter and β is the scale parameter. Thus, if is Y 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:

$$k_a = A_a/c,$$

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

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

The hot-spot hypothesis testing model is analogous to that of the binomial described previously.

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

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

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 explicitly (assuming independence).

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. Accordingly, simulation might be done by replacing unknown parameters with their maximum likelihood estimates under H_0 .

7. 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 hot-spot map depicts regions of extreme departure from expectation in the multiplicative sense, that is, 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 such as age, gender and population size before beginning the analysis (Bithell, Dutton, Neary and Vincent, 1995; Kulldorff, Feuer, Miller and Freedman, 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 hot spots 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 re-run with the adjusted sizes. Comparative configuration of new and old hot spots 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

1 functional relation is identified. Similar approaches
 2 work, in principle, for multiple factors. However, the
 3 “curse of dimensionality” often comes into play and
 4 data sparseness prevents calculation of dependable
 5 local means. Our approach, in such cases, is to cluster
 6 the data points in factor space. A mean response is then
 7 calculated for each cluster.

8
 9 **8. ILLUSTRATIVE APPLICATIONS IN ECOSYSTEM**
 10 **HEALTH AND ENVIRONMENT**

11 In this section we briefly discuss three illustrative
 12 applications in ecosystem health and environment.

13
 14 **8.1 Network Analysis of Biological Integrity in**
 15 **Freshwater Streams**

16 This study employs the network version of the upper
 17 level set scan statistic to characterize biological impair-
 18 ment along the rivers and streams of Pennsylvania and
 19 to identify subnetworks that are badly impaired. The
 20 state Department of Environmental Protection is deter-
 21 mining indices of biological integrity (IBI) at about
 22 15,000 sampling locations across the Commonwealth.
 23 Impairment is measured by a complemented form of
 24 these IBI values. Remotely sensed landscape variables
 25 and physical characteristics of the streams are used as
 26 explanatory variables to account for impairment hot
 27 spots. Critical stream subnetworks that remain unac-
 28 counted for after this filtering exercise become candi-
 29 dates for more detailed modeling and site investigation.
 30 See Evans et al. (2003), Hawkins, Norris, Hogue and
 31 Feminella (2000) and Wardrop et al. (2003).

32
 33 *Mapping of vegetation disturbance for carbon bud-*
 34 *gets.* Hot-spot detection can complement existing ap-
 35 proaches to remote measuring and mapping vegetation
 36 disturbance for global change research. Existing data
 37 products either strive to reduce “false alarms” by rely-
 38 ing on multiyear comparisons of matched “best qual-
 39 ity” data (see Strahler et al., 1999; Zhan et al., 1999,
 40 2000) or restrict information to one type of disturbance
 41 (e.g., forest fires). National and global carbon bud-
 42 gets, at time scales relevant to inversion of atmospheric
 43 transport models, require data that are both timelier
 44 and more comprehensive. Carbon management is a key
 45 area of climate change technology and, for manage-
 46 ment of carbon sequestration, vegetation disturbance
 47 needs to be detected in a manner that is timely enough
 48 both to inform management decisions and to provide
 49 feedback on the consequences of management deci-
 50 sions. [See Wofsy and Harris (2002) for an overview
 51 of existing national approaches to inventorying carbon

stocks.] The study will sample EOS data streams (pri- 52
 marily from MODIS instruments), test proposed hot- 53
 spot algorithms for their potential for support of carbon 54
 management decisions, identify data sources for hot- 55
 spot characterization (e.g., GLAS, ETM+, commercial 56
 hyperspatial) and develop ways of integrating carbon 57
 hot-spot detection and prioritization into national car- 58
 bon inventories and carbon budgets. 59

60 **8.2 Early Detection of Biological Invasions**

61
 62 Intentional and unintentional introductions of non- 62
 native exotic species have major economic and eco- 63
 logical impacts across the United States. The National 64
 Academy of Sciences estimates the cost of lost crops 65
 and containment measures at \$137 billion per year. 66
 Early detection of invasive weedy plants is the only 67
 cost-effective and tractable option for their contain- 68
 ment or eradication. However, systems for synthesiz- 69
 ing on-the-ground observation, spatial data and newly 70
 acquired remotely sensed data are lacking. We will 71
 apply the ULS scan statistic and prioritization tools 72
 to obtain more efficient surveys for invasive species 73
 and to improve the responsiveness of environmental 74
 managers to outbreaks. Japanese stiltgrass has become 75
 established in forests and waterways in the eastern 76
 United States and threatens to significantly reduce for- 77
 est and riparian species diversity and to impede wa- 78
 ter flow in rivers and streams. Often locally estab- 79
 lished populations have begun to spread before those popu- 80
 lations have been detected and likelihood of success- 81
 ful management is severely compromised. Coupling 82
 the data resources with the scan statistic represents a 83
 promising approach to preventing the transition of in- 84
 vasive plants from isolated established populations to 85
 spreading ones. See Mortensen, Johnson and Young 86
 (1993), Mortensen, Bastiaans and Sattin (2000) and 87
 Mortensen, Dieleman and Williams (2003). 88

89
 90 **ACKNOWLEDGMENTS**

91 Prepared with partial support from U.S. EPA STAR 91
 Grant for Atlantic Slope Consortium and NSF Digi- 92
 tal Government Program Grant for Geoinformatic Sur- 93
 veillance Decision Support. The contents have not been 94
 subjected to EPA review and therefore do not necessar- 95
 ily reflect the views of the agency and no official en- 96
 dorsement should be inferred. 97

98
 99 **REFERENCES**

100
 101 AARTS, E. and KORST, J. (1989). *Simulated Annealing and*
 102 *Boltzmann Machines.* Wiley, New York.

- 1 ALM, S. E. (1997). On the distribution of the scan statistic of a two-
2 dimensional Poisson process. *Adv. in Appl. Probab.* **29** 1–16.
- 3 ALM, S. E. (1998). On the distribution of scan statistics for Poisson
4 processes in two and three dimensions. *Extremes* **1** 111–126.
- 5 ANDERSON, N. H. and TITTERINGTON, D. M. (1997). Some
6 methods for investigating spatial clustering with epidemiolog-
7 ical applications. *J. Roy. Statist. Soc. Ser. A* **160** 87–105.
- 8 BITHELL, J. F., DUTTON, S. J., NEARY, N. M. and
9 VINCENT, T. J. (1995). Controlling for socioeconomic con-
10 founding using regression methods. *Community Health* **49**
11 S15–S19.
- 12 BIOMEDWARE (2001). *Software for the Environmental and Health*
13 *Sciences*. Biomedware, Ann Arbor, MI.
- 14 BRYANT, G. and MONK, P. (2001). Final report of the in-
15 vestigations into the North Leicestershire cluster of vari-
16 ant Creutzfeldt–Jakob. NHS Leicestershire Health Authority,
17 Leicestershire, UK.
- 18 CHEN, J. and GLAZ, J. (1996). Two-dimensional discrete scan
19 statistics. *Statist. Probab. Lett.* **31** 59–68.
- 20 CORMEN, T. H., LEIERSON, C. E., RIVEST, R. L. and STEIN, C.
21 (2001). *Introduction to Algorithms*, 2nd ed. MIT Press.
- 22 COUSENS, S., SMITH, P. G., WARD, H., EVERINGTON, D.,
23 KNIGHT, R. S. G., ZEIDLER, M., STEWART, G. et al. (2001).
24 Geographic distribution of variant Creutzfeldt–Jakob disease
25 in Great Britain, 1994–2000. *The Lancet* **357** 1002–1007.
- 26 CRESSIE, N. (1991). *Statistics for Spatial Data*. Wiley, New York.
- 27 D’AIGNAUX, J. H., COUSENS, S. N., DELASNERIE-
28 LAUPRETRE, N., BRANDEL, J.-P., SALOMON, D., et al.
29 (2002). Analysis of the geographical distribution of sporadic
30 Creutzfeldt–Jakob disease in France between 1992 and 1998.
31 *Internat. J. Epidemiology* **31** 490–495.
- 32 DUCZMAL, L. and ASSUNCAO, R. (2003). A simulated annealing
33 strategy for the detection of arbitrarily shaped spatial clusters.
34 *Comput. Statist. Data Anal.* To appear.
- 35 DWASS, M. (1957). Modified randomization tests for nonparamet-
36 ric hypotheses. *Ann. Math. Statist.* **28** 181–187.
- 37 EVANS, B. M., LEHNING, D. W., CORRADINI, K. J.,
38 PETERSEN, G. W., NIZEYIMANA, E., HAMLETT, J. M.,
39 ROBILLARD, P. D. and DAY, R. L. (2003). A comprehensive
40 GIS-based modeling approach for predicting nutrient loads in
41 watersheds. Unpublished manuscript.
- 42 GLAZ, J. and BALAKRISHNAN, N., eds. (1999). *Scan Statistics*
43 *and Applications*. Birkhäuser, Boston.
- 44 GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*.
45 Springer, New York.
- 46 HAWKINS, C. P., NORRIS, R. H., HOGUE, J. N. and
47 FEMINELLA, J. W. (2000). Development and evaluation of
48 predicative models for measuring the biological integrity of
49 streams. *Ecological Appl.* **10** 1456–1477.
- 50 JOHNSON, G. and PATIL, G. P. (2004). *Landscape Pattern*
51 *Analysis for Assessing Ecosystem Condition*. Kluwer, Boston.
To appear.
- KNJAZEW, D. (2002). *OmeGA: A Competent Genetic Algorithm*
for Solving Permutation and Scheduling Problems. Kluwer,
Boston.
- KNUTH, D. E. (1973). *The Art of Computer Programming I.*
Fundamental Algorithms, 2nd ed. Addison-Wesley, Reading,
MA.
- KULLDORFF, M. (1997). A spatial scan statistic. *Comm. Statist.*
Theory Methods **26** 1481–1496.
- KULLDORFF, M. (2001). Prospective time-periodic geographical
disease surveillance using a scan statistic. *J. Roy. Statist. Soc.*
Ser. A **164** 61–72.
- KULLDORFF, M., ATHAS, W. F., FEUER, E. J., MILLER, B. A.
and KEY, C. R. (1998a). Evaluating cluster alarms: A space–
time scan statistic and brain cancer in Los Alamos. *Amer.*
J. Public Health **88** 1377–1380.
- KULLDORFF, M., FEUER, E. J., MILLER, B. A. and
FREEDMAN, L. S. (1997). Breast cancer clusters in Northeast
United States: A geographic analysis. *Amer. J. Epidemiology*
146 161–170.
- 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. (1998b). *SaTScan v 2.1: Software for*
the Spatial and Space–Time Scan Statistics. National Cancer
Institute, Bethesda, MD.
- LOADER, C. R. (1991). Large-deviation approximations to the dis-
tribution of scan statistics. *Adv. in Appl. Probab.* **23** 751–771.
- MORTENSEN, D. A., BASTIAANS, L. and SATTIN, M. (2000).
The role of ecology in developing weed management systems:
An outlook. *Weed Research* **40** 49–62.
- MORTENSEN, D. A., DIELEMAN, J. A. and WILLIAMS, M. M.
(2003). Using remote sensing in integrated weed management:
What do we need to see? Unpublished manuscript.
- MORTENSEN, D. A., JOHNSON, G. A. and YOUNG, L. J. (1993).
Weed distributions in agricultural fields. In *Soil Specific Crop*
Management (P. Robert and R. H. Rust, eds.) 113–124.
Agronomy Society of America, Madison, WI.
- MOSTASHARI, F., KULLDORFF, M. and MILLER, J. (2002). Dead
bird clustering: A potential early warning system for West
Nile virus activity. New York City Department of Health, New
York, NY.
- MYERS, W. L. and PATIL, G. P. (2002). Echelon analysis. In
Encyclopedia of Environmetrics **2** 583–586. Wiley, New York.
- MYERS, W. L. and PATIL, G. P. (2004). *Doubly Segmented Images*
and Landscape Indicators for GIS Analysis: With Emphasis
on Investigation of Landscape Change. Kluwer, Boston. To
appear.
- NAUS, J. (1965a). The distribution of the size of maximum cluster
of points on the line. *J. Amer. Statist. Assoc.* **60** 532–538.
- NAUS, J. (1965b). Clustering of random points in two dimensions.
Biometrika **52** 263–267.
- PATIL, G. P. (1996). Statistical ecology, environmental statistics,
and risk assessment. In *Advances in Biometry: 50 Years of the*
International Biometric Society (P. Armitage and H. A. David,
eds.) 213–240. Wiley, New York.
- PATIL, G. P. (2002). Next generation of potential outbreak
detection and prioritization system. Invited comment and
discussion, National Syndromic Surveillance Conference,
New York City. Available at <http://www.stat.psu.edu/~gpp/PDFfiles/SyndromicSurveillance%20Comment.pdf>.
- PATIL, G. P., BISHOP, J., MYERS, W. L., TAILLIE, C.,
VRANEY, R. and WARDROP, D. H. (2002b). Detec-
tion and delineation of critical areas using echelons and
spatial scan statistics with synoptic cellular data. Techni-
cal Report 2002-0501, Center for Statistical Ecology
and Environmental Statistics, Dept. Statistics, Pennsyl-

CRITICAL AREA DETECTION, VIA SCAN STATISTICS

9

- 1 vania state Univ. Available at <http://www.stat.psu.edu/~gpp/PDFfiles/SyndromicSurveillance%20Comment.pdf>.
- 2 PATIL, G. P., BROOKS, R. P., MYERS, W. L., RAPPORT, D. J.
3 and TAILLIE, C. (2001). Ecosystem health and its measure-
4 ment at landscape scale: Towards the next generation of quan-
5 titative assessments. *Ecosystem Health* **7** 307–316.
- 6 PATIL, G. P., JOHNSON, G., MYERS, W. L. and TAILLIE, C.
7 (2000). Multiscale statistical approach to critical-area analysis
8 and modeling of watersheds and landscapes. In *Statistics for*
9 *the 21st Century: Methodologies for Applications of the Future*
10 (C. R. Rao and G. J. Szekely, eds.) 293–310. Dekker, New
11 York.
- 12 PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and
13 FLANNERY, B. P. (1992). *Numerical Recipes in C*, 2nd ed.
14 Cambridge Univ. Press.
- 15 ROGERSON, P. A. (2001). Monitoring point patterns for the
16 development of space–time clusters. *J. Roy. Statist. Soc. Ser. A*
17 **164** 87–96.
- 18 STRAHLER, A., MUCHONEY, D., BORAK, J., FRIEDL, M.,
19 GOPAL, S., LAMBIN, E. and MOODY, A. (1999).
20 MODIS land cover product, algorithm theoretical
21 basis document (ATBD), version 5.0. Available at
22 http://modis.gsfc.nasa.gov/data/atbd/atbd_mod12.pdf.
- 23 TURNBULL, B., IWANO, E. J., BURNETT, W. S., HOWE, H. L.
24 and CLARK, L. C. (1990). Monitoring for clusters of disease:
25 Application to leukemia incidence in upstate New York. *Amer.*
26 *J. Epidemiology* **132** S136–S143.
- 27 VAN EENWYK, J., BENSLEY, L., MCBRIDE, D., HOSKINS, R.,
28 SOLET, D., BROWN, A. M., TOPIWALA, H., RICHTER, A.
29 and CLARKE, R. (1999). Addressing community health con-
30 cerns around SeaTac airport. Second Report on the Work Plan
31 Proposal in August 1998, Washington State Department of
32 Health, Olympia, WA.
- 33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
- 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.)
000–000. Oxford Univ. Press.
- WALSH, S. J. and DECHELLO, L. M. (2001). Geographical
variation in mortality from systemic lupus erythematosus in
the United States. *Lupus* **10** 637–646.
- WALSH, S. J. and FENSTER, J. R. (1997). Geographical clustering
of mortality from systemic sclerosis in the Southeastern United
States, 1981–1990. *J. Rheumatology* **24** 2348–2352.
- WARDROP, D. H., BISHOP, J. A., EASTERLING, M.,
HYCHKA, K., MYERS, W. L., PATIL, G. P., and TAILLIE, C.
(2003). Use of landscape and land use parameters for classifi-
cation and characterization of watersheds in the Mid-Atlantic
across five physiographic provinces. Unpublished manuscript.
- WINKLER, G. (1995). *Image Analysis, Random Fields and Dy-*
namic Monte Carlo Methods. Springer, New York.
- WOFSY, S. C. and HARRIS, R. C. (2002). The North American
Carbon Program (NACP). Report of the NACP Committee
of the U.S. Interagency Carbon Cycle Science Program, U.S.
Global Change Research Program, Washington, DC.
- ZHAN, X., DEFRIES, R. S., HANSEN, M. C., TOWN-
SHEND, J. R. G., DIMICELI, C. M., SOHLBERG, R.
and HUANG, C. (1999). MODIS enhanced land cover
and land cover change product, algorithm theoret-
ical basis document (ATBD), version 2.0. Available at
<http://modis.gsfc.nasa.gov/data/atbd/atbd-mod29.pdf>.
- ZHAN, X., DEFRIES, R. S., TOWNSHEND, J. R. G.,
DIMICELI, C. M., HANSEN, M. C., HUANG, C. and
SOHLBERG, R. (2000). The 250m global land cover change
product from the moderate resolution imaging spectroradiome-
ter of NASA's Earth observing system. *Internat. J. Remote*
Sensing **21** 1433–1460.
- 52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102