Statistical Science
2003, Vol. 18, No. 4, 457–465

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 (2002, 2004), Patil (2002), Patil et al. (2004), 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., 2004) 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 next generation of the present day SaTScan (Kulldorff, 1997; Patil et al., 2004).

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.

2. CRITICAL AREA DETECTION WITH THE SPATIAL SCAN STATISTIC

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 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., 1998a; Kulldorff et al., 1998b; Kulldorff, 2001). 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 many issues in environmental science.

As in all geospatial surveillance, 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 the National Institutes of Health, the Centers for Disease Control and Prevention and the state health departments in New York, Connecticut, Texas, Washington, Maryland, California and New Jersey. 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 or mortality areas.

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

3. SCAN STATISTIC SUCCESS STORIES

The circular spatial scan statistic and the accompanying SaTScan software are widely used by both governmental health departments and academic epidemiologists. Some of the past and present applications include the following:

- *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, Kulldorff and Miller (2002).
- *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 preselection bias of airport area when testing the difference in the incidence rate close to the airport versus further away from the airport. A SaTScan analysis for the county as a whole revealed a nonsignificant 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 five 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).

4. PROPERTIES OF THE SCAN STATISTIC

The scan statistic is a statistical method with many potential applications, designed to detect a local excess of events and to test if such an excess can reasonably have occurred by chance. The scan statistic was first studied in detail by Naus (1965a, b), who looked at the problem in both one and two dimensions. Glaz, Naus and Wallenstein (2001) recently published a book summarizing 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 antipersonnel mines, with an interest to detect large mine fields for removal; they could be Geiger counts, with an interest to detect large uranium deposits; they could be stars or galaxies; they could be breast calcifications showing up in a mammogram, 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; 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 is quite feasible, and it is possible to analyze data sets with 10,000+ geographical locations and 100,000 cases or more.

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

Under the null hypothesis, Naus (1965b), Loader (1991) and Alm (1997, 1998) looked at a homogeneous Poisson process, Turnbull et al. (1990) considered an inhomogeneous Poisson process, and Anderson and Titterington (1997) considered both types. Chen and Glaz (1996) considered a Bernoulli model. As for the scanning window, Naus (1965b), Loader (1991), Chen and Glaz (1996), Alm (1997, 1998) and Anderson and Titterington (1997) all considered rectangles. 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.

Currently available spatial scan statistic software has several limitations. First, circles have been used for the scanning window, resulting in low power for detection of irregularly shaped clusters (Figure 1). Alternatively, an irregularly shaped cluster may be reported as a series of circular clusters. Second, the response variable has been defined on the cells of a

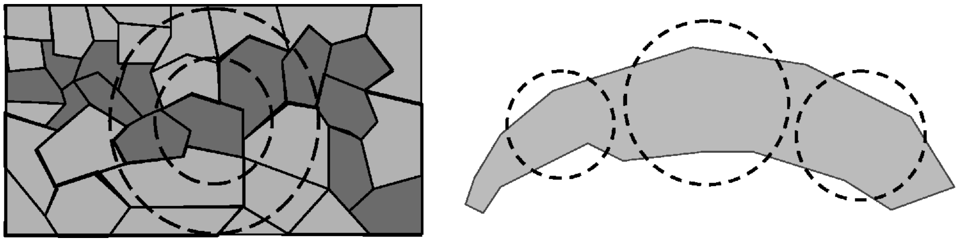

FIG. 1. *Limitations of circular scanning windows*: (left) *an irregularly shaped cluster—perhaps a cholera outbreak along a winding river floodplain*; *small circles miss much of the outbreak and large circles include many unwanted cells*; (right) *circular windows may report a single irregularly shaped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

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.

5. BASIC THEORY OF THE SCAN STATISTIC

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:

- *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 $Y_a \sim \text{Poisson}(\lambda, A_a)$, 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) 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 "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*,

$$G_a = Y_a / A_a,$$

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.

A collection of cells from the tessellation should satisfy several geometrical properties before it can be 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.

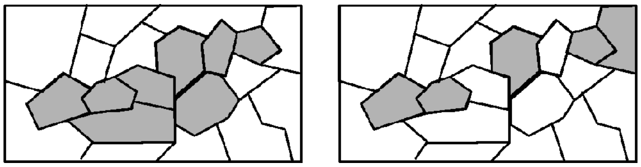

FIG. 2. *A tessellated region: the collection of shaded cells in the left diagram is connected and, therefore, constitutes a zone in Ω; the collection on the right is not connected.*

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:

$\text{H}_0 : p_a$ is the same for all cells in region R, that is, there is no hot spot.

$\text{H}_1 :$ There is a nonempty zone Z (connected union of cells) and parameter values $0 < p_0, p_1 < 1$ such that

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

The zone Z specified in H_1 is an unknown parameter of the model. The full model, $\text{H}_0 \cup \text{H}_1$, involves three unknown parameters:

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

The null model, H_0, is the limit of H_1 as $p_1 \to p_0$; 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 and 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 Assunção (2004).

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 system. 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 $a \to G_a$ 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, Leiserson, 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_α; 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 Y is a gamma-distributed variate,

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

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

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

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

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

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

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

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, 2003; 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 rerun 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 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.

8. ILLUSTRATIVE APPLICATIONS IN ECOSYSTEM HEALTH AND ENVIRONMENT

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

8.1 Network Analysis of Biological Integrity in Freshwater Streams

This study employs the network version of the upper level set scan statistic to characterize biological impairment along the rivers and streams of Pennsylvania and to identify subnetworks that are badly impaired. The state Department of Environmental Protection is determining indices of biological integrity (IBI) at about 15,000 sampling locations across the Commonwealth. Impairment is measured by a complemented form of these IBI values. Remotely sensed landscape variables and physical characteristics of the streams are used as explanatory variables to account for impairment hot spots. Critical stream subnetworks that remain unaccounted for after this filtering exercise become candidates for more detailed modeling and site investigation. See Evans et al. (2003), Hawkins, Norris, Hogue and Feminella (2000) and Wardrop et al. (2004).

8.2 Mapping of Vegetation Disturbance for Carbon Budgets

Hot-spot detection can complement existing approaches to remote measuring and mapping vegetation disturbance for global change research. Existing data products either strive to reduce "false alarms" by relying on multiyear comparisons of matched "best quality" data (see Strahler et al., 1999; Zhan et al., 1999, 2000) or restrict information to one type of disturbance (e.g., forest fires). National and global carbon budgets, at time scales relevant to inversion of atmospheric transport models, require data that are both timelier and more comprehensive. Carbon management is a key area of climate change technology and, for management of carbon sequestration, vegetation disturbance needs to be detected in a manner that is timely enough both to inform management decisions and to provide feedback on the consequences of management decisions. [See Wofsy and Harriss (2002) for an overview of existing national approaches to inventorying carbon stocks.] The study will sample EOS data streams (primarily from MODIS instruments), test proposed hot-spot algorithms for their potential for support of carbon management decisions, identify data sources for hot-spot characterization (e.g., GLAS, ETM+, commercial hyperspatial) and develop ways of integrating carbon hot-spot detection and prioritization into national carbon inventories and carbon budgets.

8.3 Early Detection of Biological Invasions

Intentional and unintentional introductions of non-native exotic species have major economic and ecological impacts across the United States. The National Academy of Sciences estimates the cost of lost crops and containment measures at $137 billion per year. Early detection of invasive weedy plants is the only cost-effective and tractable option for their containment or eradication. However, systems for synthesizing on-the-ground observation, spatial data and newly acquired remotely sensed data are lacking. We will apply the ULS scan statistic and prioritization tools to obtain more efficient surveys for invasive species and to improve the responsiveness of environmental managers to outbreaks. Japanese stiltgrass has become established in forests and waterways in the eastern United States and threatens to significantly reduce forest and riparian species diversity and to impede water flow in rivers and streams. Often locally established populations have begun to spread before those populations have been detected and likelihood of successful management is severely compromised. Coupling the data resources with the scan statistic represents a promising approach to preventing the transition of invasive plants from isolated established populations to spreading ones. See Mortensen, Johnson and Young (1993), Mortensen, Bastiaans and Sattin (2000) and Mortensen, Dieleman and Williams (2003).

ACKNOWLEDGMENTS

Prepared with partial support from U.S. EPA STAR Grant for Atlantic Slope Consortium and NSF Digital Government Program Grant for Geoinformatic Surveillance Decision Support. The contents have not been subjected to EPA review and therefore do not necessarily reflect the views of the agency and no official endorsement should be inferred.

REFERENCES

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

ALM, S. E. (1997). On the distributions of scan statistics of a two-dimensional Poisson process. *Adv. in Appl. Probab.* **29** 1–18.

ALM, S. E. (1998). Approximation and simulation of the distributions of scan statistics for Poisson processes in higher dimensions. *Extremes* **1** 111–126.

ANDERSON, N. H. and TITTERINGTON, D. M. (1997). Some methods for investigating spatial clustering with epidemiological applications. *J. Roy. Statist. Soc. Ser. A* **160** 87–105.

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

BITHELL, J. F., DUTTON, S. J., NEARY, N. M. and VINCENT, T. J. (1995). Controlling for socioeconomic confounding using regression methods. *J. Epidemiology and Community Health* **49** S15–S19.

BRYANT, G. and MONK, P. (2001). Final report of the investigation into the North Leicestershire cluster of variant Creutzfeldt–Jakob disease. NHS Leicestershire Health Authority, Leicestershire, UK. Available at http://www.cbsnews.com/htdocs/pdf/vcjd.pdf.

CHEN, J. and GLAZ, J. (1996). Two-dimensional discrete scan statistics. *Statist. Probab. Lett.* **31** 59–68.

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2001). *Introduction to Algorithms*, 2nd ed. MIT Press, Cambridge, MA.

COUSENS, S., SMITH, P. G., WARD, H., EVERINGTON, D., KNIGHT, R. S. G., ZEIDLER, M., STEWART, G. et al. (2001). Geographic distribution of variant Creutzfeldt–Jakob disease in Great Britain, 1994–2000. *The Lancet* **357** 1002–1007.

CRESSIE, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

D'AIGNAUX, J. H., COUSENS, S. N., DELASNERIE-LAUPRÊTRE, N., BRANDEL, J.-P., SALOMON, D. et al. (2002). Analysis of the geographical distribution of sporadic Creutzfeldt–Jakob disease in France between 1992 and 1998. *Internat. J. Epidemiology* **31** 490–495.

DUCZMAL, L. and ASSUNÇÃO, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Comput. Statist. Data Anal.* **45** 269–286.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Statist.* **28** 181–187.

EVANS, B. M., LEHNING, D. W., CORRADINI, K. J., PETERSEN, G. W., NIZEYIMANA, E., HAMLETT, J. M., ROBILLARD, P. D. and DAY, R. L. (2002). A comprehensive GIS-based modeling approach for predicting nutrient loads in watersheds. *J. Spatial Hydrology* **2**(2), 18 pages.

GLAZ, J. and BALAKRISHNAN, N., eds. (1999). *Scan Statistics and Applications*. Birkhäuser, Boston.

GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*. Springer, New York.

HAWKINS, C. P., NORRIS, R. H., HOGUE, J. N. and FEMINELLA, J. W. (2000). Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications* **10** 1456–1477.

JOHNSON, G. and PATIL, G. P. (2004). *Landscape Pattern 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* **1**. *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, New Mexico. *Amer. J. Public Health* **88** 1377–1380.

KULLDORFF, M., FEUER, E. J., MILLER, B. A. and FREEDMAN, L. S. (1997). Breast cancer clusters in the 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 distribution of scan statistics. *Adv. in Appl. Probab.* **23** 751–771.

MORTENSEN, D. A., BASTIAANS, L. and SATTIN, M. (2000). The role of ecology in the development of 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? *Agronomy J.* To appear.

MORTENSEN, D. A., JOHNSON, G. A. and YOUNG, L. J. (1993). Weed distributions in agricultural fields. In *Proc. Soil Specific Crop Management* (P. C. Robert, R. H. Rust and W. E. Larson, eds.) 113–124. American Society of Agronomy, 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 the maximum cluster of points on a 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. (2004). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. *Environ. Ecol. Stat.* **11**. To appear.

PATIL, G. P., BROOKS, R. P., MYERS, W. L., RAPPORT, D. J. and TAILLIE, C. (2001). Ecosystem health and its measurement at landscape scale: Toward the next generation of quantitative assessments. *Ecosystem Health* **7** 307–316.

PATIL, G. P., JOHNSON, G., MYERS, W. L. and TAILLIE, C. (2000). Multiscale statistical approach to critical-area analysis and modeling of watersheds and landscapes. In *Statistics for the 21st Century*: *Methodologies for Applications of the Future* (C. R. Rao and G. J. Székely, eds.) 293–310. Dekker, New York.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (1992). *Numerical Recipes in C*, 2nd ed. Cambridge Univ. Press.

ROGERSON, P. A. (2001). Monitoring point patterns for the development of space–time clusters. *J. Roy. Statist. Soc. Ser. A* **164** 87–96.

STRAHLER, A., MUCHONEY, D., BORAK, J., FRIEDL, M., GOPAL, S., LAMBIN, E. and MOODY, A. (1999). MODIS land cover product, algorithm theoretical basis document (ATBD), version 5.0. Available at http://modis.gsfc.nasa.gov/data/atdb/atbd_mod12. pdf.

TURNBULL, B., IWANO, E. J., BURNETT, W. S., HOWE, H. L. and CLARK, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *Amer. J. Epidemiology* **132** 136–143.

VANEENWYK, J., BENSLEY, L., MCBRIDE, D., HOSKINS, R., SOLET, D., BROWN, A. M., TOPIWALA, H., RICHTER, A. and CLARK, R. (1999). Addressing community health concerns around SeaTac airport. Second Report on the Work Plan Proposed in August 1998, Washington State Department of Health, Olympia, WA. Available at http://www.doh.wa.gov/EHSPHL/Epidemiology/NICE/publications/Seatac_Report2.pdf.

WALLER, L. (2003). Methods for detecting disease clustering in time or space. In *Monitoring the Health of Populations*: *Statistical Principles and Methods for Public Health Surveillance* (R. Brookmeyer and D. Stroup, eds.). 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. (2004). Use of landscape and land use parameters for classification and characterization of watersheds in the Mid-Atlantic across five physiographic provinces. *Environ. Ecol. Stat.* **11**. To appear.

WINKLER, G. (1995). *Image Analysis*, *Random Fields and Dynamic Monte Carlo Methods*. Springer, New York.

WOFSY, S. C. and HARRISS, R. C. (2002). The North American Carbon Program (NACP). Report of the NACP Committee of the U.S. Carbon Cycle Science Steering Group, U.S. Global Change Research Program, Washington, DC. Available at http://www.carboncyclescience.gov/nacp.pdf.

ZHAN, X., DEFRIES, R. S., HANSEN, M. C., TOWNSHEND, J. R. G., DIMICELI, C. M., SOHLBERG, R. and HUANG, C. (1999). MODIS enhanced land cover and land cover change product, algorithm theoretical basis documents (ATBD), version 2.0. Available at http://modis.umiacs.umd.edu/reports/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 250 m global land cover change product from the moderate resolution imaging spectroradiometer of NASA's Earth observing system. *Internat. J. Remote Sensing* **21** 1433–1460.