

PENNSTATE



---

## Center for Statistical Ecology and Environmental Statistics

---

### DIGITAL GOVERNANCE, HOTSPOT DETECTION, AND HOMELAND SECURITY

By G.P. Patil<sup>1</sup>, Raj Acharya<sup>2</sup>, and Shashi Phoha<sup>3</sup>

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics

<sup>2</sup>Department of Computer Science and Engineering

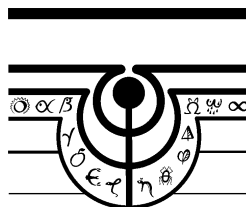
<sup>3</sup>Division of Information Science and Technology, Applied Research Laboratory  
The Pennsylvania State University, University Park, PA, 16802

<sup>3</sup>Information Technology Laboratory, National Institute of Standards and Technology,  
Gaithersburg, MD 20899

This material is based upon work supported by (1) the National Science Foundation under Grant No. 0307010, and (ii) The United States Environmental Protection Agency under Grant No. CR-83059301 and No. R-828684-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

[Invited Paper for the Encyclopedia of Quantitative Risk Analysis]

Technical Report Number 2007-0221  
TECHNICAL REPORTS AND REPRINTS SERIES  
February 2007



---

Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802

G. P. Patil  
Distinguished Professor and Director  
Tel: (814)865-9442 Fax: (814)865-1278  
Email: [gpp@stat.psu.edu](mailto:gpp@stat.psu.edu)

<http://www.stat.psu.edu/~gpp>

<http://www.stat.psu.edu/hotspots>

DGOnline News

[Environmental and Ecological Statistics-Springer](#)

# DIGITAL GOVERNANCE, HOTSPOT DETECTION, AND HOMELAND SECURITY<sup>1</sup>

by  
G.P. Patil<sup>1</sup>, Raj Acharya<sup>2</sup>, and Shashi Phoha<sup>3</sup>

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics

<sup>2</sup>Department of Computer Science and Engineering

<sup>3</sup>Division of Information Science and Technology, Applied Research Laboratory  
The Pennsylvania State University, University Park, PA, 16802

<sup>3</sup>Information Technology Laboratory, National Institute of Standards and Technology,  
Gaithersburg, MD 20899

Abstract: Effort is in progress for spatial and spatiotemporal hotspot detection, early warning, and security. A hotspot can mean an unusual phenomenon, anomaly, aberration, outbreak, elevated cluster, critical area, or object recognition and tracking. The declared need may be for monitoring, etiology, early warning, or management. The responsible factors may be natural, accidental, or intentional.

This article attempts to address concepts, methods, tools, and a variety of case studies and applications of interest to agencies, academia, and the private sector involving societal issues for homeland security, such as public health, infectious disease, invasive species, crop pathogens, biosecurity, hospital networks and syndromic surveillance, drinking water network systems, crisis management, and others.

Keywords: *Geosecurity, hotspot detection, early warning, geographic surveillance, networked infrastructure surveillance, syndromic surveillance, biosecurity, homeland security, object-recognition and tracking, upper level set scan system.*

---

<sup>1</sup> This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, and (ii) the United States Environmental Protection Agency under Grants No. CR-83059301 and No. R-828684-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

## Introduction and Background

In geospatial and spatiotemporal 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.

A declared purpose of digital governance is to empower public with information access, analysis, and policy to enable transparency, accuracy, and efficiency for societal good at large. Hotspot detection and prioritization become natural undertakings as a result of the information access over space and time. Hotspot means spot that is hot, which is of special interest or concern. Geoinformatic surveillance for spatial and temporal hotspot detection and prioritization is crucial in the 21<sup>st</sup> century. And so also the need for geoinformatic surveillance decision support system equipped with the next generation of geographic and networked hotspot detection, prioritization, and early warning with emerging hotspots.

## Scan Statistic Methodology and Technology

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, and (ii) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant. The spatial scan statistic [1] has become a popular method for detection and evaluation of disease clusters. In space-time, the scan statistic can provide early warning of disease outbreaks and can monitor their spatial spread.

**Spatial Scan Statistic Background.** The spatial scan statistic deals with the following situation. A region  $R$  of Euclidian space is tessellated or subdivided into cells that will be labeled by the symbol  $a$ . Data is available in the form of a count  $Y_a$  (non-negative integer) on each cell  $a$ . In addition, a “size” value  $A_a$  is associated with each cell  $a$ . The cell sizes  $A_a$  are regarded as known and fixed, while the cell counts  $Y_a$  are random variables. In the disease setting, the response  $Y_a$  is the number of diseased individuals within the cell and the size  $A_a$  is the total number of individuals in the cell. Generally, however, the size variable is adjusted for factors such as age, gender, environmental exposures, etc., that might affect incidence of the disease. The disease rate within the cell is the ratio  $Y_a / A_a$ . The spatial scan statistic seeks to identify “hotspots” or clusters of

cells that have an elevated rate compared with the rest of the region, and to evaluate the statistical significance ( $p$ -value) of each identified hotspot. These goals are accomplished by setting up a formal hypothesis-testing model for a hotspot. The null hypothesis asserts that there is no hotspot, i.e., that all cells have (statistically) the same rate. The

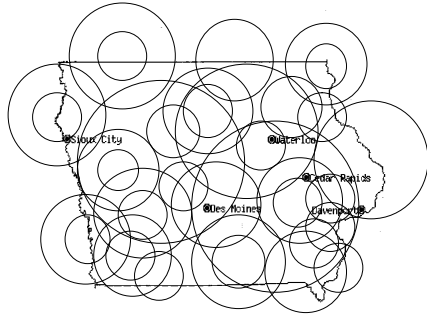


Figure 1: A small sample of the circles used

alternative states that there is a cluster  $Z$  such that the rate for cells in  $Z$  is higher than for cells outside  $Z$ . An essential point is that the cluster  $Z$  is an unknown parameter that has to be estimated. Likelihood methods are employed for both the estimation and significance testing. Candidate clusters for  $Z$  are referred to as zones. Ideally, maximization of the likelihood should search across all possible zones, but their number is generally too large for practical implementation. Various devices (e.g., expanding circles) are

employed to reduce the list of candidate zones to manageable proportions. Significance testing for the spatial scan statistic employs the likelihood ratio test; however, the standard chi-squared distribution cannot be used as reference or null distribution—in part because the zonal parameter  $Z$  is discrete. Accordingly, Monte Carlo simulation is used to determine the needed null distributions.

Explication of a likelihood function requires a distributional model (response distribution) for the response  $Y_a$  in cell  $a$ . This distribution can vary from cell to cell but in a manner that is regulated by the size variable  $A_a$ . Thus,  $A_a$  enters into the parametric structure of the response distribution. In disease surveillance, response distributions are generally taken as either binomial or Poisson, leading to comparatively simple likelihood functions.

**Limitations of Current Scan Statistic Methodology.** Available 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. 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, water distribution system, highway system, etc.). Third, response distributions have been taken as discrete (specifically, binomial or Poisson). Finally, the traditional scan statistic returns only a point estimate for the hotspot but does not attempt to assess estimation uncertainty. All of these limitations are addressed by the innovation proposed next.

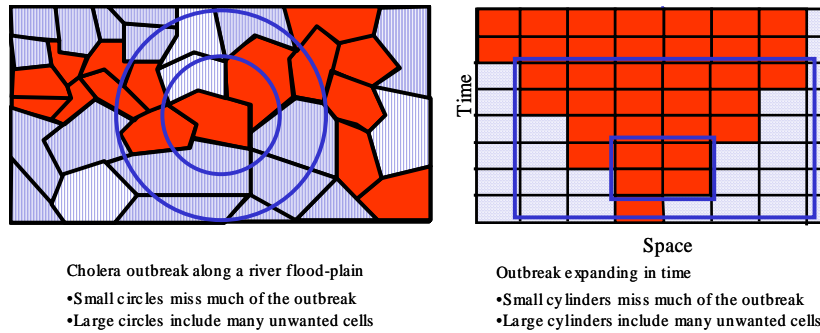


Figure 2: Scan statistic zonation for circles (*left*) and space-time cylinders (*right*).

**The Proposed Approach.** In the approach to the scan statistic, the geometric structure that carries the numerical information is an abstract graph consisting of (*i*) a finite collection of vertices and (*ii*) a finite set of edges that join certain pairs of distinct vertices. A tessellation determines such a graph: vertices are the cells of the tessellation and a pair of vertices is joined by an edge whenever the corresponding cells are adjacent. A network determines such a graph directly. Each vertex in the graph carries three items of information: (*i*) a size variable that is treated as known and non-random, (*ii*) a response variable whose value is regarded as a realization of some probability distribution, and (*iii*) the probability distribution itself, which is called the response distribution. Parameters of the response distribution may vary from vertex to vertex, but the mean response (i.e., expected value of the response distribution) should be proportional to the value of the size variable for that vertex. The response rate is the ratio Response/Size and a hotspot is a collection of vertices for which the overall response rate is unusually large.

**ULS Scan Statistic.** A new version of the spatial scan statistic is designed for detection of hotspots of arbitrary shapes and for data defined either on a tessellation or a network. This version looks for hotspots from among all connected components of upper level sets of the response rate and is therefore called the upper level set (ULS) scan statistic [2, 3]. The method is adaptive with respect to hotspot shape since candidate hotspots have their shapes determined by the data rather than by some *a priori* prescription like circles or ellipses. This data dependence will be taken into account in the Monte Carlo simulations used to determine null distributions for hypothesis testing. It will also compare performance of the ULS scanning tool with that of the traditional spatial scan statistic. The key element here is enumeration of a searchable list of candidate zones  $Z$ . A zone is, first of all, a collection of vertices from the abstract graph. Secondly, those vertices should be connected (because a geographically scattered collection of vertices would not be a reasonable candidate for a “hotspot.” Even with this connectedness limitation, the number of

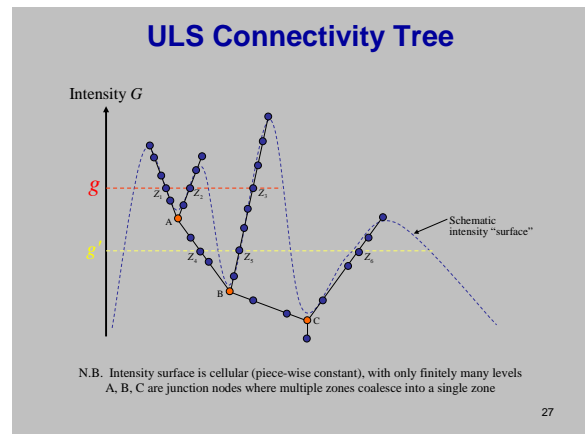


Figure 3: ULS Connectivity Tree

candidate zones is too large for a maximum likelihood search in all but the smallest of graphs. The list of zones is reduced to a searchable size in the following way. The response rate at vertex  $a$  is  $G_a = Y_a / A_a$ . These rates determine a function  $a \rightarrow G_a$  defined over the vertices in the graph. This function has only finitely many values (called levels) and each level  $g$  determines an upper level set  $U_g$  defined by  $\{a : G_a > g\}$ . Upper level sets do not have to be connected but each upper level set can be decomposed into the disjoint union of connected components. The list of candidate zones  $Z$  for the ULS scan statistic consists of all connected components of all upper level sets. This list of candidate zones is denoted by  $\Omega_{\text{ULS}}$ . The zones in  $\Omega_{\text{ULS}}$  are certainly plausible as potential hotspots since they are portions of upper level sets. Their number is small enough for practical maximum likelihood search—in fact, the size of  $\Omega_{\text{ULS}}$  does not exceed the number of vertices in the abstract graph (e.g., the number of cells in the tessellation). Finally,  $\Omega_{\text{ULS}}$  becomes a tree under set inclusion, thus facilitating computer representation. 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; the root node consists of all vertices in the abstract graph.

### 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 employs cylindrical extensions of spatial circles and cannot detect the temporal evolution of a hotspot. The space-time generalization of the ULS scan detects arbitrarily shaped hotspots in space-time. This helps classify space-time hotspots into various evolutionary types—a few of which appear on the left hand side of the following figure. The merging hotspot is particularly interesting because, while it comprises a connected zone in space-time, several of its time slices are spatially disconnected.

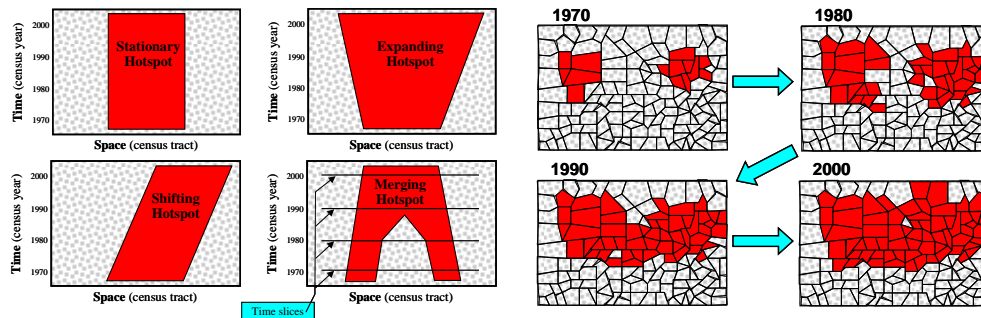


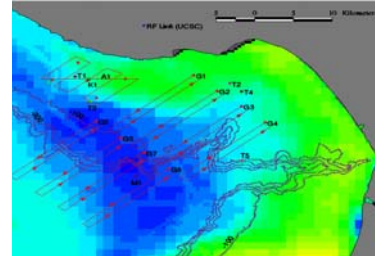
Figure 4: The four diagrams on the left depict different types of space-time hotspots. The spatial dimension is shown schematically on the horizontal and time is on the vertical. The diagrams on the right show the trajectory (sequence of time slices) of a merging hotspot.

### KEY APPLICATIONS

Broadly speaking, the proposed geosurveillance project and its forum identify several case studies around the world. This section provides a selection of illustrative applications and case studies.

**Tasking of a self-organizing oceanic surveillance mobile sensor network.**

The Autonomous Ocean Sampling Network Simulator (AOSN) is used to study coordination and control strategies for high-resolution, spatio-temporally coordinated surveys of oceanographic fields such as bathymetry, temperature, and currents using autonomous unmanned undersea vehicles. Currently, the network of mobile sensor platforms is autonomous and self-organizing once given high-level tasking from an external tactical coordinator. This case study proposes to use upper level set scan statistic theory to identify hotspots in data gathered by the sensor network and use this information to dynamically task mobile sensor platforms so that more data can be gathered in the areas of interest. By detecting hotspots and tasking accordingly, network resources are not wasted on mapping areas of little change. The ability of the sensor network to dynamically task its own components expands the network’s mission and increases the reactivity to changing conditions in a highly dynamic environment. [17, 18, 19]



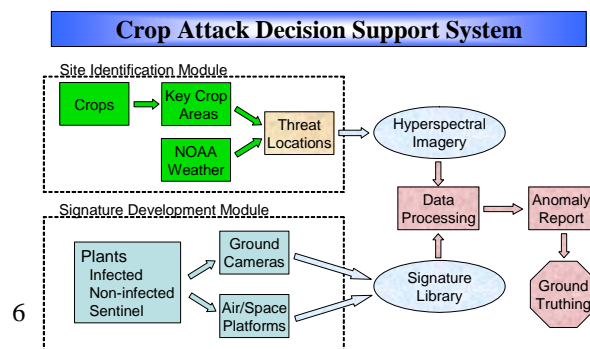
**Early detection of biological invasions.**

Intentional and unintentional introductions of non-native exotic species have major economic and ecological impacts across the US. A National Academy of Sciences report [7] 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. But 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 US and threatens to significantly reduce forest and riparian species diversity, and impede water flow in rivers and streams. Often locally established populations have begun to spread before those populations have been detected and likelihood for 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, like that depicted in the photograph. [8, 9]



**Development of remote sensing methods for crop bioterrorism.**

Testimony to the House National Security Subcommittee this past fall stressed the high probability that crop terrorism could and would occur.



The logistics of ground-based monitoring are daunting, and point to a strong need to monitor US cropland by remote sensing. Remote sensing, to be timely and effective, should resolve zones of infection as small as 5m in diameter. Improvements in the next generation of sensors will have to be made, as will the techniques for handling the huge volumes of high resolution data. Another difficulty has been development of high-quality signatures detectable from airborne or space-borne platforms. To-date, reliable detectable signatures have been difficult to obtain. We propose to defeat this impasse by integrating biologists, engineers, and statisticians to generate high quality hyperspectral signatures, to determine signal strength on unique portions of these signatures, and to develop goals for instrument resolution from various platforms. Specific goals are: (i) Develop signatures for two key pathogens using ground-based, portable hyperspectral cameras. (ii) Determine signal strength vs. noise of plants infected with single or multiple pathogens and/or insects. (iii) Enhance hardware and processing algorithms to filter and resolve crop pathogen signals. (iv) Provide goals for the next generation of air and space-borne sensors for high-threat pathogens. The diagram depicts the process from signature development through identification of threat areas, signal acquisition/processing to the development of an anomaly report. [10, 11]

### **Cyber Security and Computer Network Diagnostics**

Securing the nation's computer networks from cyber attacks is an important aspect of national Homeland Security. Network diagnostic tools aim at detecting security attacks on computer networks. Besides cyber security, these tools can also be used to diagnose other anomalies such as infrastructure failures, and operational aberrations. Hotspot detection forms an important and integral part of these diagnostic tools for discovering correlated anomalies. It is constructive to develop a network diagnostic tool at a functional level. The goal of network state models is to obtain the temporal characteristics of network elements such as routers, typically in terms of their physical connectivity, resource availability, occupancy distribution, blocking probability, etc. There is some prior work [12] in developing network state models for connectivity, and resource availability. Models have been also developed for studying the equilibrium behavior of multi-dimensional loss systems. The probabilistic finite state automaton (PFSA) describing a network element can be obtained from the output of these state models. A time-dependent crisis-index is determined for each network element, which measures their normal behavior pattern compared to crisis behavior. The crisis-index is the numerical distance between the stochastic languages generated by the normal and crisis automata. Use of the variational distance between probability measures seems attractive, although other distances can also be considered. The crisis behavior can be obtained from past experience. The crisis indices over a collection of network elements are then used for hot-spot detection using scan statistic methodology. These hot spots help to detect coordinated security attacks geographically spread over a network.

### **Drinking Water Quality and Water Utility Vulnerability**

New York City has installed 892 drinking water sampling stations across the five boroughs. Each 4.5-foot high station is located outdoors and draws water from a nearby water main. The purpose is to monitor general



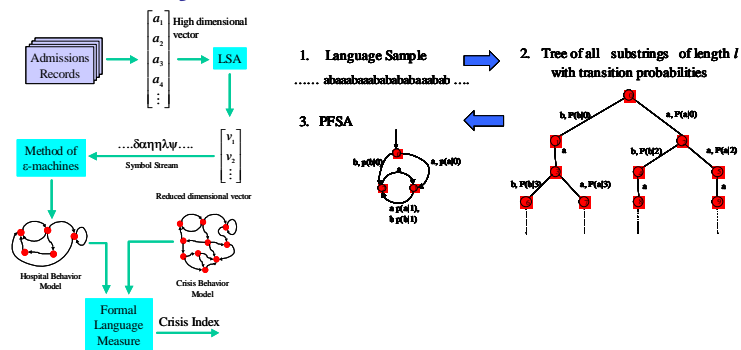
water quality, detect potential health threats, and thwart bioterror activity. Sampling frequency was increased after the 9/11 attacks and, currently, about 47,000 water samples are analyzed annually. Parameters analyzed include: bacteria, chlorine, pH, inorganic and organic pollutants, color, turbidity, odor, and many others. The network version of the ULS scan statistic can provide a real-time surveillance system for detecting and evaluating water quality hotspots within the distribution system on a parameter-by-parameter basis.

An overall assessment of water quality at each sampling station taking all parameters into account is achieved by employing recent progress on multi-criterion ranking using poset (partially ordered set) prioritization [13].

### Syndromic Surveillance Network and Early Warning

Emerging hotspots for disease, biological agents, or medical effects of pollution are identified through modeling events at local hospitals. A time-dependent crisis index is determined for each hospital in a network spread over a city, state or the whole country. This index measures the behavior patterns at each hospital compared to crisis behavior. The behaviors are based on series of hospital admission records containing symptoms and diagnoses. The basic components of behaviors are events, which in this case are hospital admissions. The important attributes of admissions are the information on the admission records and how frequently admissions are occurring compared to normal, non-crisis behavior. The behavior stream is represented as probabilistic finite state automaton and by the corresponding formal stochastic language. The method of epsilon-machines [14, 15] is used to estimate the automaton from the current behavior stream. The variational distance between stochastic languages provides a quantitative measure of how close the current behavior is to that of a crisis. This distance is called the crisis index. The crisis index over the network of hospitals is used for hotspot detection by the upper level set scan statistic.

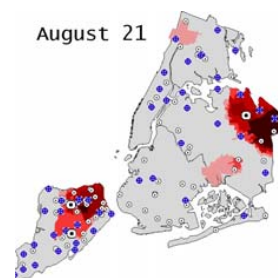
### Syndromic Surveillance



(left) The overall procedure, leading from admissions records to the crisis index for a hospital. The hotspot detection algorithm is then applied to the crisis index values defined over the hospital network. (right) The  $\epsilon$ -machine procedure for converting an event stream into a parse tree and finally into a probabilistic finite state automaton (PFSA).

### West Nile Virus: An Illustration of the Early Warning Capability of the Scan Statistic

Since the 1999 West Nile (WN) virus outbreak in New York City, health officials have been searching for an early warning system that could signal increased risk of human WN infection, and provide a basis for targeted public education and increased



mosquito control. Birds and mosquitoes with laboratory evidence of WN virus preceded most human infections in 2000, but sample collection and laboratory testing are time-consuming and costly. The cylinder-based space-time scan statistic for detecting small area clustering of dead bird reports have been assessed for its utility in providing an early warning of WN virus activity [16].

All unique non-pigeon dead bird reports were categorized as 'cases' if occurring in the prior 7 days and 'controls' if occurring during a historic baseline. The most likely cluster area was determined using the scan statistic and its statistical significance was evaluated using Monte Carlo hypothesis testing. Analyses were performed in a prospective simulation for 2000 and in real-time during 2001.

For the 2000 data, dead bird clustering was found in Staten Island over 4 weeks prior to laboratory evidence of WN virus in birds and mosquitoes from this area. Real-time implementation in 2001 led to intensified larval control in eastern Queens, over three weeks prior to laboratory confirmation from this cluster area. Dead bird clusters were identified around the residence of five of seven human infections, from 0 to 40 days (median 12) prior to the onset of illness, and 12-45 days (median 17) prior to human diagnosis.

It has been concluded that scan statistical cluster analysis of dead bird reports may provide early warning of viral activity among birds and of subsequent human infections. Analysis with the ULS space-time scan statistic will be worthwhile. Since the latter allows for arbitrarily shaped clusters in both the spatial and temporal dimensions, there is potential for earlier detection with a reduced false alarm rate.

### **New York City Subway System Syndromic Surveillance**

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

The New York City Department of Health (DOH) and Metropolitan Transportation Authority (MTA) began monitoring subway worker absenteeism in October 2001 as one of several surveillance systems for the early detection of disease outbreaks. Each day the MTA transmits an electronic line list of workers absent the previous day, including work location and reason for absence. DOH epidemiologists currently monitor temporal trends in absences in key syndrome categories (e.g., fever-flu or gastrointestinal illness). Analytic techniques are needed for detecting hotspots within the subway network.

## **Looking Forward**

We now live in the age of geospatial technologies. The age old geographic issues of societal importance are now thinkable and analyzable. The geography of disease is now just as doable as genetics of disease, for example. And it is possible to pursue it in an intelligent manner with the rapidly advancing information technology around.

Government agencies continue to require meaningful summaries of georeferenced data to support policies and decisions involving geographic assessments and resource allocations. So also the public with initiatives of digital governance in the country and around the world.

This article briefly describes a prototype hotspot detection system for hotspot delineation and provides a variety of case studies and illustrative applications of societal importance for homeland security.

Surveillance geoinformatics of hotspot detection and prioritization is a critical need of the 21<sup>st</sup> century. Next generation decision support system within this context is crucial. It will be productive to build on the present effort in the directions of prototype and user-friendly methods, tools, and software, and also in the directions of thematic groups, working groups, and case studies important at various scales and levels. The authors have such a continuation effort in progress within the context of digital governance with NSF support and would welcome interested readers to join in this collaborative initiative.

## References

- [1] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26, 1481–1496.
- [2] Patil, G.P. & Taillie, C. (2004a). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11, 183-197.
- [3] Patil, G.P., Duczmal, L., Haran, M., & Patankar, P. (2006a). On PULSE: The progressive upper level set scan statistic system for geospatial and spatiotemporal hotspot detection. In: The 7<sup>th</sup> Annual International Conference on Digital Government Research, San Diego, CA, May 21-24, 2006.
- [4] Phoha, S., Peluso, E., Stadter, P., Stover, J., & Gibson, R. (1997). A mobile distributed network of autonomous undersea vehicles, Proceedings of the 24<sup>th</sup> Annual Symposium and Exhibition of the Association of Unmanned Vehicle Systems Intl., Baltimore, MD.
- [5] Phoha, S., Eberbach, E., & Brooks, R. (1999). Coordination of multiple heterogeneous autonomous undersea vehicles (AUVs). Special Heterogeneous Multi-Robot Systems Issue of *Autonomous Robots*, October 1999.
- [6] Phoha, S., Peluso, E., & Culver, R.L. (2001). A high fidelity ocean sampling mobile network (SAMON) simulator tested for evaluating intelligent control of unmanned underwater vehicles. Revised version submitted to *IEEE Journal for Ocean Engineering*, Special Edition, June 2001.
- [7] National Academy of Sciences. (2002). *Predicting Invasions of Nonindigenous Plants and Plant Pests*. National Academy Press, Washington, D.C.
- [8] Mortensen, D.A., Bastiaans, L., & Sattin, M. (2000). The role of ecology in developing weed management systems: an outlook. *Weed Research*, 40, 49–62.
- [9] Mortensen, D.A., Dieleman, J.A., & Williams, M.M. (2002). Using remote sensing in integrated weed management: what do we need to see? *Agronomy Journal* (in press).
- [10] Backman, P.A. & Jacobi, J.C. (1997). Developing thresholds for plant disease management. Pages 114–127 in *Economic Thresholds for Integrated Pest Management*, L. Higley and L. Pedigo, eds. University of Nebraska Press.
- [11] Wheelis, M., Casagrande, R., & Madden, L.V. (2002). Biological attack on agriculture: Low-tech high-impact bioterrorism. *BioScience*, 52, 569–576.
- [12] Ghosh, D. & Acharya, R. (2001). A probabilistic scheme for hierarchical routing. *Proceedings of Ninth IEEE International Conference on Networks*. IEEE. pp. 416-421.

- [13] Patil, G.P. & Taillie, C. (2004b). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics*, 11, 199-228.
- [14] Shalizi, C.R., Shalizi, K.L., & Crutchfield, J.P. Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. *Journal of Machine Learning Research*. (2002) submitted.
- [15] Shalizi, K.L., Shalizi, C.R., & Crutchfield, J.P. Pattern discovery in time series, Part II: Implementation, evaluation, and comparison. *Journal of Machine Learning Research*. (2002) to be submitted.
- [16] Mostashari, F., Kulldorff, M., Hartman, J., Miller, J., & Kulasekera, V. (2003). Dead bird clusters as an early warning system for West Nile Virus activity. *Emerging Infectious Diseases*, 9(6), 641-646.