

**GEOINFORMATIC SURVEILLANCE FOR HOTSPOT DETECTION AND PRIORITIZATION  
Innovation with Epsilon Machines, Formal Language Measures, Upper Level Set Scans,  
Partially Ordered Set Prioritizations, Decision Support Systems,  
and Virtual Situation Room Servers**

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

**GEOINFORMATIC SURVEILLANCE FOR HOTSPOT DETECTION AND PRIORITIZATION  
Innovation with Epsilon Machines, Formal Language Measures, Upper Level Set Scans,  
Partially Ordered Set Prioritizations, Decision Support Systems,  
and Virtual Situation Room Servers**

## Table of Contents

Abstract	2
Project Description	3
1. Introduction and Motivation	3
2. Fundamental Information Technology Research and Its Novel Application	4
3. Illustrative Applications and Case Studies	5
4. Upper Level Set Scan Statistic Methodology and Technology	9
5. Partially Ordered Set Prioritization Methodology and Technology	12
6. Geoinformatic Surveillance Decision Support System	14
7. Research Training, Education, and Technology Transfer	16
8. Project Outcomes, Partner Synergy, and Workplan	17
9. Management Plan and Structure	18
10. References	19

### Abstract

Current methods to organize, represent, and process large bodies of complex information spread over space and time are inadequate for today's decision making needs, especially in a time of crisis. Advances are needed in methods of quickly and accurately recognizing and prioritizing critical changes in important parameters that are masked by fluctuations. We propose research that will address these needs in crisis situations, as well as the non-crisis infrastructure needs of science and technology that are equally important for interpreting high-dimensional multi-attribute spatio-temporal information for policy and research.

Our project will conduct fundamental information science and technology research and its novel application to geoinformatic surveillance for hotspot detection and prioritization. A hotspot means something unusual—an anomaly, aberration, outbreak, elevated cluster, critical area, etc. The declared need may be for monitoring, etiology, management, or early warning. Responsible factors may be natural, accidental, or intentional.

The most innovative aspect of this research develops *upper level set scan statistic theory* to recognize arbitrarily-shaped hotspots. Spatio-temporal data are integrated with a new level of accuracy providing more sensitive indicators of changes in critical parameters. The technique applies not only to physical space, but also to connected collections of objects or regions, i.e. networks. A second innovation is the development of *partially ordered set prioritization theory* to rank hotspots without having to integrate multiple indicators into a single index. A third is a new method of automated knowledge acquisition in the form of *behavior recognition* technology built on the concept of  $\varepsilon$ -complexity and  $\varepsilon$ -machines from Statistical Physics and a *formal language measure* from Discrete Event Control Theory.

Our research consists of three parts. First, fundamentally new information technologies are developed from advances in statistics, statistical physics and control theory. A new level of sensitivity is attained for recognizing and responding to critical changes in noisy, chaotic environments. Second, the technological advances are proven in test cases covering a broad range of critical situations. The range of applications demonstrates the fundamental nature of the new technologies. Third, we will move our advances into society by building prototype *situation room servers*. The servers will integrate complex distributed data sources for selected

applications. These servers and the new tools they make available will revolutionize crisis prediction and management.

Toward the end of the grant, we will find interested agencies and make the technology available in an ongoing, operational capacity. This will ensure that the benefits of our research will have a long-term impact on society. The project will also have a strong educational component with effective technology transfer, outreach, and built-in evaluation.

Keywords: biosecurity, carbon budget, computer network diagnostics, crop pathogens, cyber security, disease surveillance, early warning system, ecosystem health, environmental justice, epsilon-machine, hotspot, hotspot rating, inter-disciplinary and interagency activities, invasive species, middleware, mobile sensor network, multicriterion prioritization, public health, syndromic surveillance, upper level set scan statistic, virtual situation room, water management.

## **Project Description**

### **1. Introduction and Motivation**

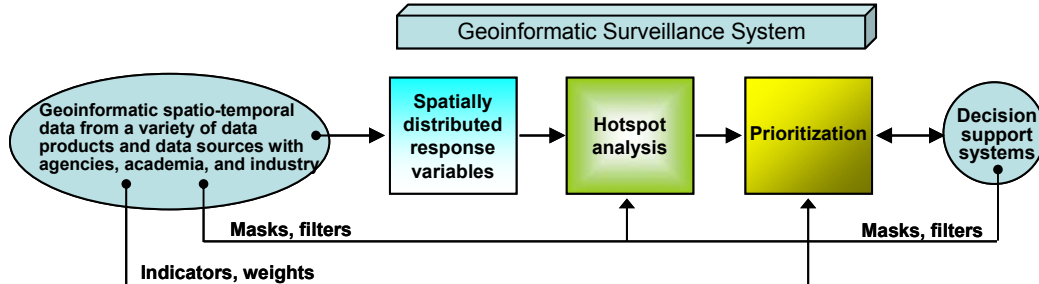
A primary purpose of this proposal is to invent, implement, and interface innovative information technology (IT) for the much needed geoinformatic surveillance decision support system for hotspot detection and prioritization in the project-networked virtual situation room capable of online interaction, cross-cutting solution, and dynamically updated communication with application partners, educational users or decision makers involved in a real situation.

Geographic surveillance for hotspot detection and delineation has become an important area of investigation both in geospatial ecosystem studies and in geospatial public health studies. In order to find critical areas based on synoptic cellular data, geospatial ecosystem investigations applied recently discovered echelon tools (Myers et al 1997, 1999). In order to find elevated rate areas based on synoptic cellular data, geospatial public health investigations apply recently discovered SaTScan, circle-based spatial scan statistic tool (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995). The PI (Patil, 2003; Patil, Balbus et al. 2003; Patil, Bishop et al., 2002, 2003) has conceptualized a joint role for these together in the spirit of a cross-disciplinary cross-fertilization to accomplish more effective and efficient geographical surveillance for hotspot detection, and early warning system.

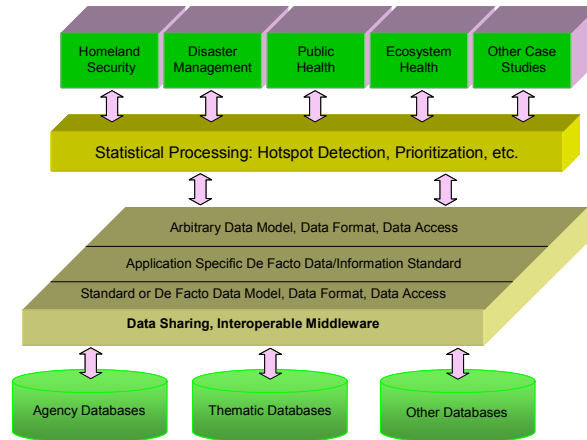
Clearly, clusters are clusters. They can be of any shape, and cannot be captured only by circles. This is likely to give more false alarms and more false negatives than warranted. What we need is the capability to detect arbitrarily shaped clusters and the ability to handle network-based as well as cell-based data. The upper level set scan system innovation will fill this need and provide a timely next generation hotspot detection and delineation system (Patil, 2002; Patil, Myers et al 2002; Myers, Kurihara et al 2002, 2003; Patil, Brooks et al 2001, 2002). Also see Patil, Balbus et al. (2003) for a broad perspective of multiscale advanced raster map analysis system of which hotspot detection is a part.

The significance and the timeliness become clearer as we witness various reports and action plans of various federal, state, and local agencies and prestigious foundations and academies, suggesting geographic and network surveillance for arbitrarily shaped hotspots, using next generation of sophisticated hotspot detection and prioritization tools. For example, a recent NRC report on making the nation safer: the role of science and technology in countering terrorism.

While the proposed research derives its particular significance within the context of national homeland security, it has powerful place within the much broader infrastructure of science and technology. Major information flows in the geoinformatic surveillance can be represented schematically as follows:



The case studies in this project address a broad range of national applications such as homeland security, biosecurity, disaster management, public health, ecosystem health, water management, carbon budget, coastal management, community infrastructure, etc. The geographic information sharing middleware will provide the component to support distributed, dynamic data-driven applications and case studies, and enhance the system security and stability. This middleware will access appropriate databases for supporting the case-studies (see Figure below).



Our team involves researchers with solid track records in several complementary areas that are at the core of this project. We will integrate the resulting advances into a prototype system applied to a rich set of large-scale case studies. Project goals and results will be achieved in a well-integrated disciplinary and cross-disciplinary effort coupled with matching educational abilities, leading to an emergent software system. It will help strengthen the methodology and technology infrastructure needed nationally for hotspot detection and prioritization across geographic regions and across networks in the 21<sup>st</sup> century, and provide a basic foundation to an envisioned National Center serving society’s need for geoinformatic surveillance.

## 2. Fundamental Information Technology Research and Its Novel Application

Current methods to organize, represent, and process large bodies of complex information spread over space and time are inadequate for today’s decision making needs, especially in a time of crisis. Advances are needed in methods of quickly and accurately recognizing and prioritizing

critical changes in important parameters that are masked by fluctuations. We propose research that will address these needs in crisis situations, as well as the non-crisis infrastructure needs of science and technology. Our project will conduct fundamental information science and technology research and its novel application to geoinformatic surveillance for hotspot detection and prioritization. A hotspot means something unusual—an anomaly, aberration, outbreak, elevated cluster, critical area, etc. The declared need may be for monitoring, etiology, management, or early warning. Responsible factors may be natural, accidental, or intentional.

The most innovative aspect of this research develops *upper level set scan statistic theory* to recognize arbitrarily shaped hotspots. Spatio-temporal data are integrated with a new level of accuracy providing more sensitive indicators of changes in critical parameters. The technique applies not only to physical space, but also to connected collections of objects or regions, i.e. networks. A second innovation is the development of partially ordered set prioritization theory to rank hotspots without having to integrate multiple indicators into a single index. A third is a new method of automated knowledge acquisition in the form of *behavior recognition* technology built on the concept of  $\varepsilon$ -complexity and  $\varepsilon$ -machines from Statistical Physics and a *formal language measure* from Discrete Event Control Theory. Local behaviors can now be compared to known behaviors using traditional pattern-matching techniques for classification. Behaviors are represented symbolically by formal languages in a form that can be used directly for automated decision aides in the form of discrete event controllers.

Our research consists of three parts. First, fundamentally new information technologies are developed from advances in statistics, statistical physics and control theory. A new level of sensitivity is attained for recognizing and responding to critical changes in noisy, chaotic environments. Second, the technological advances are proven in test cases covering a broad range of critical situations. The range of applications demonstrates the fundamental nature of the new technologies. Third, we will move our advances into society by building prototype *situation room servers*. The servers will integrate complex distributed data sources for selected applications. These servers and the new tools they make available will revolutionize crisis prediction and management.

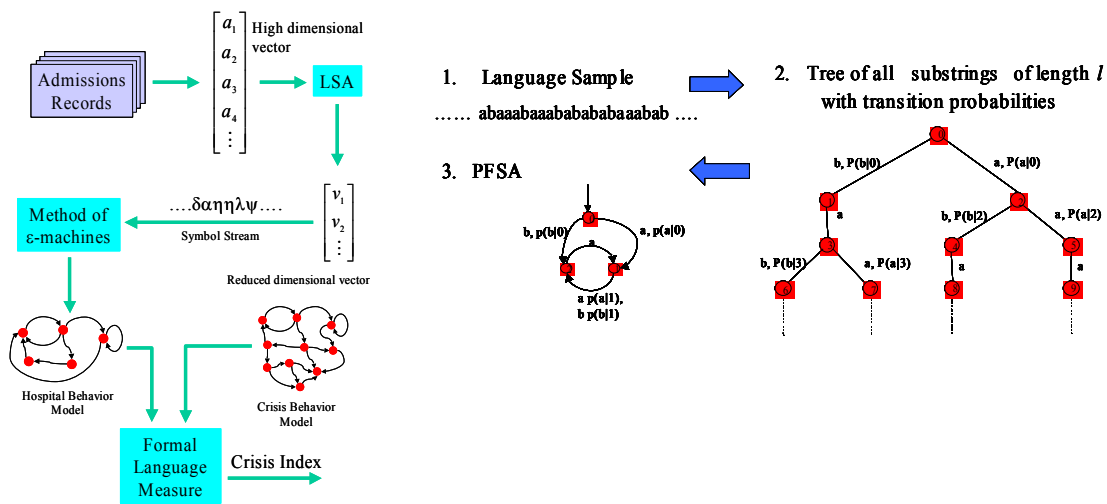
Toward the end of the grant, we will find interested agencies and make the technology available in an ongoing, operational capacity. This will ensure that the benefits of our research will have a long-term impact on society. The project will also have a strong educational component with technology transfer, outreach, and built-in evaluation.

### 3. Illustrative Applications and Case Studies

Broadly speaking, the proposed geosurveillance project identifies several case studies important for the national applications. In this section, we present five illustrative applications and case studies, with a view to provide the feel. The supplement of this proposal (Section I) contains 35 description summaries of 16 prototype case studies (PCS), 13 user case studies (UCS), and 6 international case studies (ICS).

**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 various symptoms and diagnoses, reported in response to various federal, state, and local informational

network programs for disease surveillance, syndromic surveillance, etc. (Friedlander et al., 2002; Kulldorff et al., 2002; Patil, 2002). Two recent breakthroughs in information science allow us to represent behaviors as *formal languages* and determine a quantitative measure of how close the current behavior is to that of a crisis. The first is from ongoing research at the Santa Fe Institute (Crutchfield, 1989, 1994; Shalizi, 2002ab) that resulted in the method of  $\epsilon$ -machines, which can construct probabilistic finite-state automata (PFSA) from a stream of symbolic events. The second is from ongoing research at Penn State (Ray, 2002; Wang, 2002) that resulted in a *formal language measure*, which can determine a quantitative distance between two behaviors represented as PFSA. We have also done research<sup>1</sup> in using the measure to determine the behaviors of robots from observations made by distributed sensor networks (Friedlander, 2000, 2003) and created a *behavior recognition tool*. We will extend this concept to hospital admissions behaviors. We briefly describe each step of this process (see Figure 1).



**Figure 1. (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).**

The basic components of behaviors are *events*, which in our 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. In current systems they are typically assigned to one of a small number of predetermined classes. Our research will refine this procedure by representing each admissions event using Latent Semantic Analysis (LSA) (Deerwester et al., 1990).

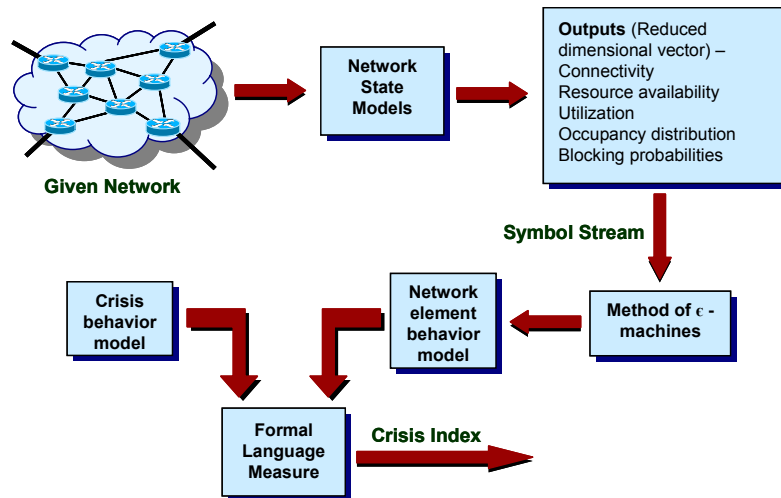
Another innovative aspect of this research is that we look at *behaviors* rather than individual events. A behavior is defined by the different sequences of events that occur at the hospital. If we take a sample of the symbol stream, the behavior is represented by all of the substrings in the sample up to some length  $l$ . The behavior representation is therefore a formal language over the admission events' alphabet. The method of  $\epsilon$ -machines models the behavior associated with an event stream by creating a probabilistic finite state machine that can recognize the substrings in

<sup>1</sup> Emergent Surveillance Plexus MURI Award No. DAAD19-01-1-0504 sponsored by the Defense Advance Research Projects Agency (DARPA), and administered by the Army Research Office.

the event stream. These machines provide a symbolic representation of the behavior in question.

Finally, we look at a formal language measure (Ray, 2002; Wang, 2002) that can be used to determine the quantitative distance between two formal languages. The measure gives us the ability to use traditional pattern matching techniques on abstract behaviors. In our case, we use the formal languages derived from the admission records of known crises as exemplars. They are matched against the current behavior to derive a *crisis index*. The crisis index over the network of hospitals is used for hotspot detection.

**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. The proposed research will be used to develop a network diagnostic tool, as shown in Figure 2 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. We have done prior work (Ghosh and Acharya, 2001; Sarangan et al., 2001, 2002) in developing network state models for connectivity, and resource availability. We have also developed models for studying the equilibrium behavior of multi-dimensional loss systems (Acharya, 2003). The 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 behavior can be obtained from past experience. The crisis indices over a collection of network elements are then used for hot-spot detection. These hot spots help to detect coordinated security attacks geographically spread over a network.



**Figure 2. Procedure for obtaining the crisis index for network elements.**

**Tasking of Self-Organizing Surveillance Mobile Sensor Networks.** Many critical applications of surveillance sensor networks involve finding hotspots. The proposed *upper level set scan statistic system* will be used to guide the search by estimating the location of hotspots based on the data previously taken by the surveillance network (Phoha et al., 2002). As mobile

sensor platforms move toward estimated hotspot locations, more data will be taken and used to update estimated hotspot locations. There are many important area surveillance applications for the proposed research including:

- Finding hotspots for radioactivity and chemical or biological agents to prevent or mitigate the effects of terrorist attacks or to detect nuclear testing.
- Mapping elevation or wind, and bathymetry or ocean currents to better understand and protect the environment.
- Detecting emerging failures in a complex networked system like the electric grid
- Mapping the gravitational field to find underground chambers or tunnels for rescue or combat missions.

Mobile sensor platforms can measure data fields along their trajectories. We are interested in using feedback from individual sensor platforms, communicated to other platforms in the network (Eberbach, 1999), to guide the search. Once measurements have been taken and communicated, the hotspot locations will be estimated using upper level set scan statistics. This information will be used to modify the search. Additional measurements will then be taken and the feedback process will repeat until the goal is reached. There are two types of hotspots in the applications listed above. The first is caused by point sources such as radioactive material. The second is interesting distributed features, e.g. an area of variability of the field that is being mapped, e.g. elevation, bathymetry or pressure. By detecting only the significant variations, resources are not wasted on mapping areas of little change.

**Oil Spill Detection, Monitoring, and Prioritization.** Damage produced by marine oil spills includes soiled beaches, bird and mammal mortality, destruction of fisheries, impaired recreational facilities, and catastrophic impairment to entire ecosystems. Remote sensing can be used for oil spill detection and prevention of further damage. For example, the Exxon Valdez slick was detected through SPOT satellite data, the Ixtoc I well blowout slick in Mexico was detected using GOES and AVHRR on the NOAA polar series satellites, and oiled ice on Gabarus Bay (Kurdistan) was detected using LANDSAT data. We will use hyperspectral image analysis of Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Synthetic Aperture Radar (SAR) data to conduct case studies of the Patuxent River in Maryland and the Santa Barbara shoreline of California for oil spill detection on sea water and associated mitigation. The main objective of using AVIRIS is to identify, measure, and monitor constituents of the Earth's surface and atmosphere based on molecular absorption and particle scattering signatures. SAR's ability to penetrate cloud cover, to illuminate the Earth's surface with its own signal, and to precisely measure distances, makes it especially useful for detecting and monitoring oil spills. The project's scan statistic hotspot delineation and poset prioritization tools will be used in combination with our oil spill detection algorithm to provide for early warning and spatial-temporal monitoring of marine oil spills and their consequences (Fingas, 1991; Kafatos and Chi, 2002; Salem and Kafatos, 2001).

**Investigating Emerging Environmental Issues of Ecosystem Health.** Many potentially critical environmental issues are indefinite in their early stages, even if comprehensively mapped on a global basis using advanced satellite remote sensing technology. Possible progressive environmental effects of global warming and associated issues of emissions and carbon management are prime examples of this indefiniteness during onset. The basic question is how severe and spatially consistent do occurrences need to be in order to constitute pattern as opposed to background levels of long-term and essentially random fluctuation? Degradation of water quality across stream networks, spread of non-native invasive organisms, and build-up of

toxic substances in soils or estuarine substrates are further examples of environmental concerns that show patterns progressively expressing over long scales of time and large scales of space (Brooks et al., 2002; Knox, 2002; Mortensen and Rathbun, 2002; Wardrop et al., 2002). Earthwatch is necessary in such contexts, but the sentinel must have an objective means of raising an alarm and pointing to sectors of larger extents that are the most likely to be exhibiting onset of problem conditions. New information must be interpreted in the context of prior information so that trends can be detected. This calls for dynamic formal statistical inference that is spatially and temporally cognizant (Myers et al., 1997, 1999, 2003). Contemporary SaTScan methodology has several limitations that lead to tentativeness of inferences regarding emergent phenomena. SaTScan currently uses spatial geometries such as circles that are often patently inappropriate to the process of concern such as influences on hierarchically convergent stream systems. The treatment of progression in time is likewise overly simplistic in relation to the possible types of trends. The proposed research addresses these shortcomings of contemporary information technology as reflected in prospective case studies including:

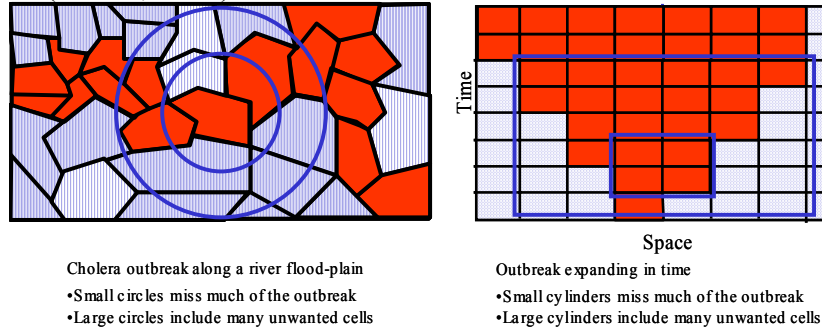
- Network analysis of biological integrity in freshwater streams
- Watershed prioritization for impairment and vulnerability
- Mapping priority hotspots of vegetative disturbance for carbon budgets
- Early detection and delineation of outbreaks of invasive plant species

#### 4. Upper Level Set 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, (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; Waller, 2003) has become a popular method for detection of disease clusters. In space-time, the scan statistic can provide early warning of disease outbreaks and can monitor their spatial spread. With innovative modifications, scan statistic methods can be used for hotspot analysis in any field. We will develop methodology and corresponding software for applications of the scan statistic to critical areas of national concern in the 21<sup>st</sup> century.

**Spatial Scan Statistic Background.** The spatial scan statistic deals with the following situation. A region  $R$  of Euclidian space is tessellated into cells that will be labeled by the symbol  $a$ . Data is available in the form of a response value  $Y_a$  on each cell  $a$ . The spatial scan statistic seeks to identify “hotspots” or clusters of cells that have an elevated response 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 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 null distribution—in part because the zonal parameter space is finite. Accordingly, Monte Carlo simulation (Dwass, 1957) is used to determine the needed null distributions.



**Figure 3. Scan statistic zonation for circles (left) and space-time cylinders (right).**

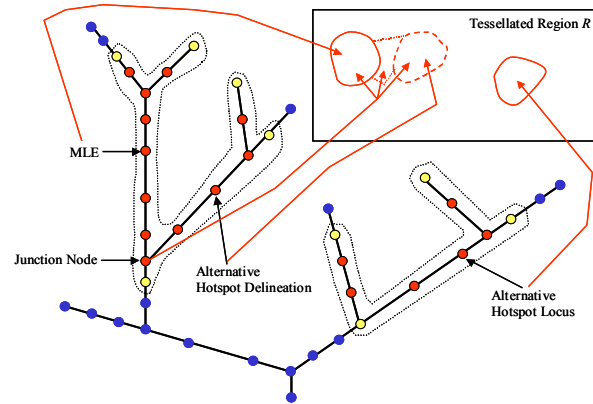
**Limitations of Current Scan Statistic Methodology.** Available scan statistic software suffers from several limitations. First, circles have been used for the scanning window, resulting in low power for detection of irregularly shaped clusters (Figure 3). Second, the response variable has been defined 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. We address all of these limitations.

**Upper Level Set Scan Statistic.** We will develop a new version of the spatial scan statistic designed for detection of hotspots of arbitrary shapes and for data defined either on a tessellation or a network. Our 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. 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.

The list of candidate zones  $Z$  for the ULS scan statistic consists of all connected components of all upper level sets. This list 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 cells in the tessellation or nodes in the network. Finally,  $\Omega_{\text{ULS}}$  becomes a tree under set inclusion, thus facilitating computer representation. This tree is called the **ULS-tree** (Figure 4); its nodes are the zones  $Z \in \Omega_{\text{ULS}}$ . Leaf nodes are (typically) singleton vertices at which the response rate is a local maximum; the root node consists of all cells in the network or all nodes in the network.

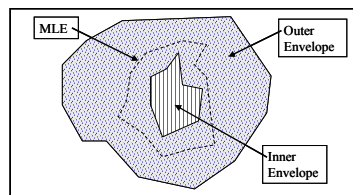
Finding the connected components for an upper level set is essentially the issue of determining the transitive closure of the adjacency relation defined by the edges of the graph. Several generic algorithms are available in the computer science literature (Cormen et al, 2001, Section 22.3 for depth first search; Knuth, 1973, p. 353 or Press et al., 1992, Section 8.6 for transitive closure).

**Continuous Response Distributions.** In extending the scan statistic methodology to include continuous responses, we will focus on three parametric families of distributions: gamma distribution, lognormal distribution, and scaled beta distribution. The first two families apply to responses that can range from zero to infinity, while the third is for bounded responses.



**Figure 4. A confidence set of hotspots on the ULS tree. The different connected components correspond to different hotspot loci while the nodes within a connected component correspond to different delineations of that hotspot.**

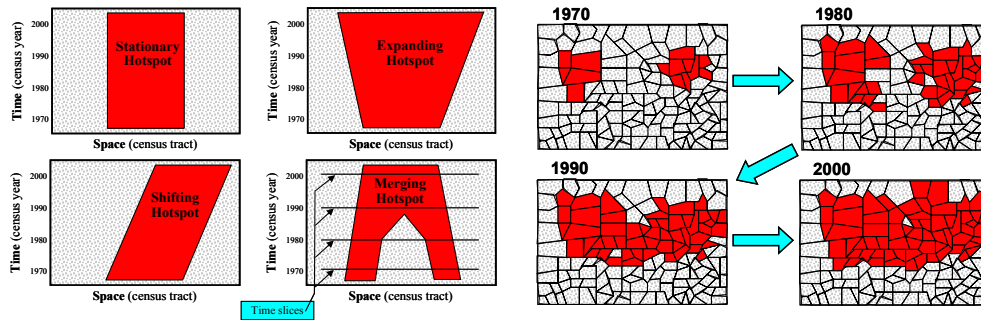
**Hotspot Confidence Sets.** The hotspot MLE is just that—an estimate. Removing some cells from the MLE and replacing them with certain other cells can generate an estimate that is almost as plausible in the likelihood sense. We will express this uncertainty in hotspot delineation by a confidence set of hotspot zones—a subset of the ULS tree (Figure 4). We will determine the confidence set by employing the standard duality between confidence sets and hypothesis testing (Lehmann, 1986, p. 90, 214) in conjunction with the likelihood ratio test. The confidence set also lets us assign a numerical **hotspot-membership rating** to each cell (e.g., county, zip code, census tract). The rating is the percentage of zones (in the confidence set) that include the cell under consideration (Figure 5). A map of these ratings, with superimposed MLE, provides a visual display of uncertainty in hotspot delineation.



**Figure 5. Hotspot-membership rating. Cells in the inner envelope belong to all plausible estimates (at specified confidence level); cells in the outer envelope belong to at least one plausible estimate. The MLE is nested between the two envelopes.**

**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 (Figure 3). The space-time generalization of the ULS scan detects arbitrarily shaped hotspots in space-time. This lets us classify space-time hotspots into various evolutionary types—a few of which appear

on the left hand side of Figure 6. 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 6.** 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.

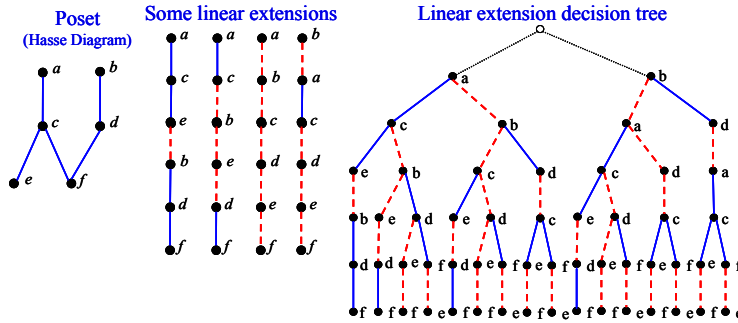
### 5. Partially Ordered Set Prioritization Methodology and Technology

We address the question of ranking a collection of objects, such as initial hotspots, when a suite of indicator values is available for each member of the collection (Patil and Taillie, 2002, 2003). The objects can be represented as a cloud of points in indicator space (Filar and Ross, 2001), but the different indicators typically convey different comparative messages and there is no unique way to rank the objects. A conventional solution assigns a composite numerical score to each object by combining the indicator information in some fashion. Every such composite involves judgments (often arbitrary or controversial) about tradeoffs among indicators. We take the view that the relative positions in indicator space determine only a partial ordering (Fishburn, 1985; Neggers and Kim, 1998; Trotter, 1992) and that a given pair of objects may not be inherently comparable. Working with Hasse diagrams (Neggers and Kim, 1998; Di Battista et al., 1999) of the partial order, we will study the collection of all rankings that are compatible with the partial order, and derive the prioritization and ranking consistent with the data matrix of objects and indicators.

**Multiple Indicators and Partially Ordered Sets (Posets).** The scan statistic ranks hotspots based on their statistical significance (likelihood values). But, other factors need to be considered in prioritizing hotspots, such as mean response, peak response, geographical extent, population size, economic value, etc. We therefore envision a suite of indicator values attached to each hotspot with large indicator values signifying greater hotspot importance. Different indicators reflect different criteria and may rank the hotspots differently. In mathematical terms, the suite of indicators determines a partial order on the set of hotspots. There are many different ways of ranking the hotspots while remaining consistent with the partial ordering.

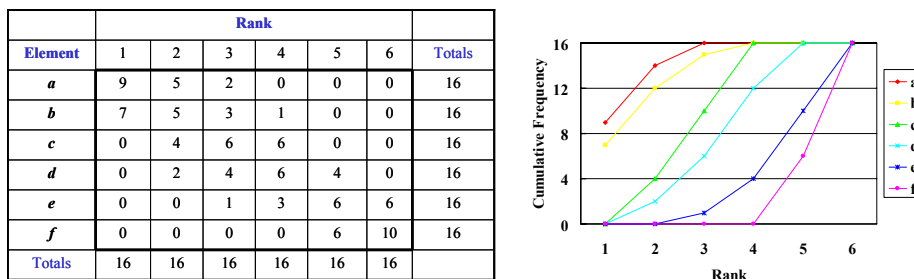
**Hasse Diagrams and Linear Extensions.** Posets can be displayed as Hasse diagrams (Figure 7). A Hasse diagram is a graph whose vertices are the hotspots and whose edges join vertices that cover one another in the partial order. Hotspot  $b$  is said to *cover*  $a$  in the partial order if three things happen: (i)  $a < b$ ; (ii)  $a \neq b$ ; and (iii) if  $a < x < b$  then either  $x = a$  or  $x = b$ . In words,  $b$  is strictly above  $a$  and no hotspots are strictly between  $a$  and  $b$ . Each of the many possible ways of ranking the elements of a poset is referred to as a linear extension. The Hasse diagram of each linear extension appears as a vertical graph (Figure 7). Enumeration of all possible linear extensions can be accomplished algorithmically as follows. The top element of a

linear extension can be any one of the maximal elements of the Hasse diagram. Select any one of these maximal elements and remove it from the Hasse diagram. The second ranked element in the linear extension can be any maximal element from the reduced Hasse diagram. Select any of these and proceed iteratively. The procedure can be arranged as a decision tree (Figure 7) and each path through the tree from root node to leaf node determines one linear extension.

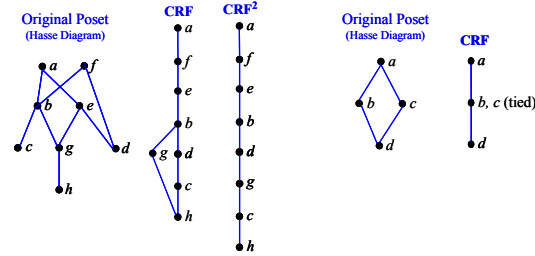


**Figure 7.** Hasse diagram of a hypothetical poset (*left*), some linear extensions (*middle*), and a decision tree giving all 16 possible linear extensions (*right*). Links shown in dashed/red (called jumps) are not implied by the partial order. The six members of the poset can be arranged in  $6!=720$  different ways, but only 16 of these are valid linear extensions.

**Linearizing a Poset.** The suite of indicators determines only a partial order on the hotspots, but it is human nature to ask for a linear ordering. We ask: Is there some objective way of mapping the partial order into a linear one? Our solution treats each linear extension in Figure 7 as a voter and we apply the principle of majority rule. Focus attention on some member  $a$  of the poset and ask how many of the voters give  $a$  a rank of 1? 2? Etc. The results are displayed in Figure 8, where each row of the table is called a rank-frequency distribution. The cumulative forms of these rank-frequency distributions form a new poset with stochastic ordering of distributions as the order relation. For this example, the new poset is already a linear ordering (see Figure 8).



**Figure 8.** (*Left*) Rank-frequency table for the poset of Figure 7. Each row gives the number of linear extensions that assign a given rank  $r$  to the corresponding member of the poset. Each row is referred to as a rank-frequency distribution. (*Right*) Cumulative rank-frequency distributions for the poset of Figure 7. The curves are stacked one above the other giving a linear ordering of the elements:  $a > b > c > d > e > f$ .



**Figure 9. (Left) Two iterations of the CRF operator are required to transform this partial order into a linear order. (Right) A poset for which the CRF operator produces ties.**

We refer to the above procedure as the cumulative rank-frequency (**CRF**) operator. In general, it does not transform a partial order into a linear order in a single step; instead, multiple iterations may be required (Figure 9). The CRF operator can also produce ties in the final linear ordering.

Enhancement of the prioritization tools will address the following needs:

- **Markov Chain Monte Carlo (MCMC) Sampling.** Except for very small posets, it is computationally impossible to enumerate all the linear extensions. For instance, a recent UNEP HEI poset has 141 members but the number of linear extensions exceeds  $8 \times 10^{105}$ . Instead of full enumeration, we will use MCMC methods to *estimate* the (row-normalized) rank-frequency table. This entails sampling from the uniform distribution on the set  $\Omega$  of all linear extensions of a given poset. See Aldous (1987), Brightwell and Winkler (1991) and Haggstrom (2002) for elaboration of MCMC methods applied to discrete data structures.
- **Measurement and Estimation Error.** Indicators are subject to measurement error, which can affect ranking results. Similarly, rank-frequencies are estimated in the MCMC version of the CRF operator whose results are therefore subject to estimation error. We will use multiple comparison and fuzzy comparison methods to assess consequences of such errors.
- **Non-uniform (Weighted) Distributions.** The CRF operator treats each linear extension as an equal “voter” in arriving at a final ranking. It is sometimes preferable to weight certain linear extensions more heavily if, for example, a particular indicator is especially important. This induces a weighted distribution on  $\Omega$  and MCMC methods will be used to sample from  $\Omega$  according to this non-uniform distribution.

## 6. Geoinformatic Surveillance Decision Support System

**Online Decision Support System:** This component of the project focuses on the development of software system to assist scientists to apply the hotspot detection and prioritization technology and application users to use the system. With dramatic technology advances in online storage and internet networking, more and more data are immediately and easily available and accessible. This enables the development of online data services in the data system, in which data system processes the user’s request and delivers the results to the user directly. A key function of the system is to enable service providers to define and register a data processing service in the system and their users to use the service to act upon the data represented in the system. It is also important to develop the standard of data process development, and the communication protocol between the system and the data service.

The project will use this software system to collect and manage the information of data and data processes. The software system has user interfaces of Web, API, and query language. We will develop the web-based user interface to provide wide interactive access to the system and timely update the dynamical information in the underlying database. We will also use and further

develop the software system to enable scientists to define and register their hotspot detection and prioritization data processes as data services in the system, and to enable the application users to use the system in support of their decisions in response to the hotspots. This will form an online Geoinformatic surveillance decision support system.

In the data system, we will create database tables to record data usage and data service activities to track system usage for performance metrics measurement. The system will also record user comments and satisfaction levels for our decision support system.

**Computational Structure, System Integration, and Database Management:** This component of the project focuses on the development of efficient data structures and algorithms coupled with efficient visualization techniques for hotspot detection and prioritization using upper level set scans and partially ordered set prioritizations. In fact, recently the problem has been for quickly identifying regions for multivariate maps for which a number of geospatial parameters satisfy certain conditions. See JaJa and Shi (2001), K.-S. Yang et al. (2001), and R. Yang et al. (2001). This project will extend these techniques in directions dictated by the proposed scanning techniques and prioritization tools.

**Information Visualization, User Interface Design and GIS Linkage:** A key goal of this effort is to develop a visualization interface integrated with software tools based on various statistical models and techniques developed in this project. Information visualization and interface design are critical to making effective use of the various models and techniques. Our goal will be to promote the discovery of inherent structures and patterns, build and test hypotheses, enable the detailed study of particular facets and dimensions of the data, and provide means to visually assess the utility and accuracy of the statistical and computational techniques developed. Our approach will be to work with applications partners to identify their needs and frequent tasks. A phased implementation will allow us to implement simple algorithms at first and then embed more sophisticated algorithms. In conjunction with case-study validations and initial demonstrations, we will conduct usability tests to refine the interfaces and demonstrate efficacy.

**Online Information System:** This component of the project focuses on the development of an information system for communication among the partners and for communication with the general users. We will develop the project web site, which will describe significant information products involving the project and its progress. It will provide central access point to locate such information, and identify opportunities to provide comments and other feedbacks during the development of some of these products. The main capabilities of the information system include:

- Tight interactions with operational users over several phases, e.g., project formulation, defining the products and services, testing-validation-refinement, and technology transfer. On-going community involvement supported through a web interface for communicating about hotspots and an interactive system for knowledge sharing and discovery.
- Support for spatial, temporal, and/or subject matter queries of hotspot data products, associated metadata, comments, and documents.
- Solicitation and archiving of responses, comments, etc., tagged with links to the relevant products, locations, documents, and/or time points (product date(s), as well as the usual header information for the communication).
- Development of a path for feedback about hotspot data and priority rankings to flow back to how indicators are ranked and combined.
- Evaluation of the hotspot interface using different views, depending on application area and role in the project (investigator, application partner, education partner, student, resident/citizen of a particular

geographical region, etc.).

**Online Virtual Situation Room:** Public and educational involvement in evaluating hotspot products will be a vital complement to more formal or structured feedback from application partners. The information system will be accessible not only by project partners and target application users but also widely distributed general public users. This information system will form the virtual situation room for distributed partners and general users benefit from the timely interaction and communication involving the real life situations at their end.

## 7. Research Training, Education, and Technology Transfer

**Rationale.** Geographic surveillance for hotspot detection, delineation, and prioritization has become an important national need in science, technology, and society. The project based cross cutting solution is a frontier solution to detect and prioritize arbitrarily shaped hotspots, regardless of whether the base data and information are cell-based or network-based and whether they are in space, in time, or in space-time. The cross cutting solution has capability to detect emerging hotspots for early warning and direct the concerned parties to investigational areas.

**Goals.** The project education drive will be to create a strong and flexible architecture that enables scalable, systemic, and sustainable solutions for the national needs in geographic hotspot detection and prioritization and to support a new national capability that enables continuous, engaging and dynamic learning with a built in provision for ongoing evaluation and assessment of and from end users of the geographic surveillance system using geospatial data sources and informational products.

**Audience.** We have the following audiences targeted:

- (a) Graduate students engaged in research involving geospatial data, hotspot detection and prioritization.
- (b) Graduate faculty and researchers involved with graduate students in (a)
- (c) Informal learning of all interested and concerned citizens and decision makers interested in and/or involved with geosurveillance activity based on synoptic and/or network-based data.

The investigators strongly believe in the broadening of opportunities to enable the participation of all citizens, with a special commitment to the principle of diversity. We will make a special effort to seek the participation of under-represented groups in our research programs.

**Graduate Education and Technology Transfer.** An essential part of this project is to introduce the concepts and the methods at the core of the system to researchers in various resource management communities. This effort will be greatly facilitated by the inclusion of prominent researchers from these communities on our team as we have planned and the availability of the system that will provide an easy way to apply these methods to various domains.

In graduate education, the investigators will integrate project technologies into a range of related graduate courses offered by the three universities involved. Also, graduate students supported on this project will be expected to contribute to the tutorials offered during each summer workshop, in addition to presenting their research progress. Every effort will be made to iteratively accomplish the upward spiral of horizontal and vertical research and training integration.

For effective technology transfer, we will develop two monographs and two casebooks covering the techniques and tools and the applications case studies. The material for these monographs and casebooks will be developed over the years, and tested incrementally through graduate and training courses offered by the investigators. The graduate students will be heavily involved in this process, and also the internet. Special issues of relevant journals will also be considered for

publication. These products and outcomes will help evaluate the success of the educational activities of the project.

**Dissemination and Evaluation.** Dissemination and evaluation will be an integral part of decision support tools and system for knowledge management. In addition, the project's education working group will meet annually to evaluate progress and assess the educational impact of the project.

## 8. Project Outcomes, Partner Synergy, and Workplan

A sixty-month duration of October 1, 2003 to September 30, 2008 is planned. Demonstrations will lead to operational use for identified case studies. User organization funds will be sought for demonstrations and scaling studies of user organization case studies, and for new applications.

### Year 1:

1. Set up a web site with up to date information on algorithms, software, and case studies, and a discussion forum that will include feedback from users of related software.
2. Introduce analytical concepts and methods to partners and develop workplans for case studies. Develop education plan and post-project implementation plans for prototype case studies.
3. Formulate upper level set scan statistic algorithms and MCMC prioritization algorithms.
4. Develop  $\varepsilon$ -machine and crisis index algorithms for hospital admissions and syndromic surveillance.

### Year 2:

1. Develop and implement modeling and simulation algorithms for scanning and prioritization systems, including CRF operator. Use in graduate classes and student research.
2. Train case study teams in concepts and methods of analysis at the core of the detection and prioritization system, while applying techniques to case studies.
3. Acquire initial samples of hospital records for NYC and Pennsylvania networks.
4. Evaluate  $\varepsilon$ -machine and crisis index algorithms for hospital admissions and syndromic surveillance.
5. Evaluate data sources and begin simulation studies for hotspot-tasking of mobile sensor networks.

### Year 3:

1. High performance algorithms and software for region analysis that includes hotspot surveillance and multi-criteria decision support. Test and validate on case studies.
2. Evaluate scanning and prioritization tools using case studies and educational activities.
3. Validate  $\varepsilon$ -machine and crisis index algorithms for hospital admissions and syndromic surveillance.
4. Design operational version of syndromic surveillance system for NYC and PA hospital networks.
5. Validate hotspot feedback algorithms for tasking of mobile sensor networks.
6. Functional decision support system for the overall project.

### Year 4:

1. Package and disseminate scanning and prioritization technology, and solicit feedback.
2. Design of situation room server for syndromic surveillance and tasking of mobile sensor networks.
3. Demonstrate decision support application using the prototype case studies.
4. Refinement of scanning/prioritization tools in conjunction with prototype case studies.
5. Formulate lessons-learned monographs on project-based scanning methods, prioritization tools, and casebooks based on case studies and educational experiences.

### Year 5

1. Demonstration of scanning and prioritization system in conjunction with applications case studies.
2. Operational support of situation room server for syndromic surveillance and mobile sensor networks.
3. Two monographs on scanning methods and on prioritization methods.
4. Casebooks summarizing Prototype Case Studies and User Organization Case Studies.
5. Operational project-wide decision support system and virtual situation room.
6. Workshop in Washington DC to report on major techniques, tools, and applications experience.

7. International Workshop in Parma, Italy with the broad-based Map of Italian Nature as the host.

## 9. Management Plan and Structure

The overall management of the project will be the responsibility of the PI, Dr. G. P. Patil. He brings a substantial administrative experience in managing crossdisciplinary research initiatives and large projects. The management team of the project will consist of the PI and the co-PIs who will set short-term and long-term directions and goals, implementation plans, assess project progress, and establish collaborative mechanisms among the participating investigators. The management and research team enjoys the kind of cross-disciplinarity the current solicitation would like to see. The goals stated in the section on Project Outcomes, Partner Synergy, and Workplan will provide initial guidance for the management team. The management team will be in constant communication through email and phone calls, and will meet twice a year in conjunction with the two annual workshops planned for the project. These meetings will focus on assessing progress, adjusting goals and directions as appropriate, and setting new goals.

Moreover, each member of the management team will lead the coordination of the research efforts in a thrust area as follows:

- Geoinformatic Surveillance – Dr. Patil
- Information Science and Technology – Dr. Phoha, Dr. Acharya, and Dr. Yen
- Data Mining, Information Fusion, and Visualization – Dr. Acharya
- Upper Level Set Detection System – Dr. Taillie
- Partially Ordered Set Prioritization System – Dr. Taillie
- Public Health Applications – Dr. Kulldorff
- Environmental and Ecological Applications – Dr. Myers
- Decision Support System – Dr. Kafatos
- Virtual Situation Room – Dr. Kafatos, Dr. Patil, and Dr. Phoha  
 Disaster and Dispersion: Kafatos; Syndromic Surveillance: Friedlander; Sensor Networks: Phoha; Disease Clusters: Kulldorff; Ecosystem Critical Areas: Wardrop; Poset Prioritization: Taillie
- Education, Dissemination, and Outreach – Dr. Patil

The two planned workshops, one in the middle of the academic year and the other in the summer, will be an integral part of the project, which will allow investigators to describe their work and how it relates to the overall goals of the project, and to assist the investigators to better integrate their expertise to help evolve the upper level set scan statistic system and the partially ordered set prioritization system with prototype case studies, and the national and international refinement and validation case studies. The summer workshop will include special tutorial sessions covering the system and the methodologies.

Most of the proposed case studies have been carefully selected as ongoing funded projects whose successful completion would significantly benefit from application of the proposed toolkit for hotspot detection and prioritization. Thus, the sponsor will have served as catalytic sponsor of methodology and technology directed toward current and future opportunities and challenges.

The proposed project has versatile partnership. It also has versatile leadership. It will have Theory and Practice Advisory Council (TAPAC) consisting of Science Advisory Board (SAB) and User Advisory Council (UAC). It will ensure that science and user interest are represented throughout the project. It will also be a strong source of relevant domain expertise. It is

fortunate to have fitting triple leadership with John Kelmelis, Chief Scientist for Geography at USGS as the TAPAC Chair; Mike Goodchild, National Academy Member and Director of the NSF Center for Spatially Integrated Social Science as SAB Chair; and Chris Portier, Director of the National Environmental Toxicology Program at NIH as UAC Chair. The TAPAC will meet once a year during the summer workshop, evaluate project progress, and advise the management team on future direction.

Partner federal agencies include CDC, DOD, DOT, EPA, NASA, NIH, NOAA, USDA, and USGS with USGS as the Coordinating Agency. The TAPAC will meet within the first ninety days of the Project to help initiate a Partners Plan for suitable funding and strategic goals for the envisioned National Center for Geosurveillance upon Project completion. The Project will play the role of embryonic National Center during its fourth and fifth year duration. The Partners Plan will help match the requested project budget with costshare and kindshare direct with PI, Co-PI, Co-Investigators, and/or Case Study Scientists. Agency needs for this type of toolkit are sufficiently pressing, and we will look forward to full operational adoption, once its efficacy has been demonstrated. Based on the feedback we have, we are optimistic. We also expect partnership and leadership to grow to serve this timely mission.

Finally, the concept of mobility and interactive visits will be fully explored and implemented between participating faculty, graduate students, and postdocs across the three universities. Each university group will have weekly miniseminar(s) on relevant themes of the project involving local faculty, graduate students, and post docs. These will be carefully strengthened from time to time with visiting collaborators from participating institutions to keep the individual and collective momentum and synergy in progressive development. Every effort will be made to iteratively accomplish the upward spiral of horizontal and vertical research and training integration.

**International Collaboration.** This project will have an impressive international collaboration with scientific leaders responsible for major geospatial landscape level spatio-temporal programs in their countries such as Brazil, Canada, India, Italy, Japan, and Sweden. The international synergies and benefits to be gained from these collaborations include the considerable enrichment of our case studies that will strengthen the empirical and validation aspects of our project. The PI has already given a series of training courses to the Italian Map of Nature and will be giving a start off workshop in Sweden in the summer to help initiate their case studies as an integral part of the Swedish Program of Remote Sensing for the Environment. The Supplement to this proposal (Section I) contains support letters and describes the international collaborations.

**Performance Metrics and Built-in Evaluation.** Beyond reporting project inputs, outputs, and the number and kinds of applications supported, and their status, we will develop ways of measuring and reporting the individual project outcomes and the impact of the project. This will be done in a timely manner on an annual basis.

## 10. References

Acharya R (2003). Steady state probabilities for multi-dimensional loss networks with bursty arrivals (under preparation).

- Aldous D (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Probability in the Engineering and Informational Sciences*, 1, 33–46.
- Brightwell G and Winkler P (1991). Counting linear extensions. *Order*, 8, 225–242.
- Brooks R, Myers WL, Patil GP, and Taillie C (2002). Prioritization model for diagnosis of watershed impairment and vulnerability. PCS-2, this proposal, pp. I-10–16.
- Colwell, R (2003). Obstinate issues, sophisticated solutions: Environmental science and education for a new age. John H. Chafee Memorial Lecture, Third National Conference on Science, Policy, and the Environment. Washington DC.
- Cormen TH, Leieron CE, Rivest RL, and Stein C (2001). *Introduction to Algorithms*, Second Edition. MIT Press, Cambridge, Massachusetts.
- Cressie N (1991). *Statistics for Spatial Data*. Wiley, New York.
- Crutchfield JP and Young K (1989). Inferring statistical complexity. *Physical Review Letters*, 63, 105–108.
- Crutchfield JP (1994). The Calculi of Emergence: Computation, Dynamics, and Induction, *Physica D*, special issue on the Proceedings of the Oji International Seminar on Complex Systems—from Complex Dynamics to Artificial Reality, April 1993, Numazu, Japan.
- Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), 391–407.
- Di Battista G, Eades P, Tamassia R, and Tollis IG (1999). *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, Upper Saddle River, New Jersey.
- Dwass M (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181–187.
- Eberbach E and Phoha S (1999). SAMON: Communication, cooperation, and learning of mobile autonomous robotic agents, Proceedings of the 11<sup>th</sup> IEEE Intl. Conf. On Tools with Artificial Intelligence, Chicago, IL.
- Filar JA and Ross NP (2001). Generalized data envelopment analysis, and environmental indicators. Invited Paper. Plenary Forum on Environmental Indicators and their integration for Quality of Life. Index 2001 Congress, Rome, Italy.
- Fingas M (1991). The technology of oil spill remote sensing. U.S. Coast Guard Oil Spill Remote Sensing Workshop, ERIM Report #213259.
- Fishburn PC (1985). *Interval Orders and Interval Graphs: A Study of Partially Ordered Sets*. Wiley, New York.
- Friedlander D, Patil GP, Phoha S, and Taillie C (2002). Emerging hotspot detection through analysis of networked patient records for crisis prediction. PCS-9, this proposal, pp. I-48–53.
- Friedlander D, Phoha S, and Brooks RR (2003). Determination of vehicle behavior based on distributed sensor network data. Submitted to SPIE International Symposium on Optical Science and Technology, 3–8 August 2003.

- Friedlander D, Ray A, and Phoha S (2000). Domain independent measures of intelligent control. *NIST's Performance Metrics for Intelligent Systems Workshop*, Gaithersburg, MD, August 14–16, 2000.
- Ghosh D and Acharya R (2001). A probabilistic approach to hierarchical QoS routing. *Ninth IEEE International Conference on Networks (ICON)*, Bangkok, October 2001.
- Haggstrom O (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, Cambridge.
- JaJa J and Shi Q (2001). *Efficient Techniques for Exploring Geospatial Data*. (submitted for publication).
- Kafatos M and Chi Y (2002). Oil spill hotspot detection and prioritization. PCS-7, this proposal, pp. I-28–29.
- Knox R (2002). Prototyping simplified methods of mapping priority hotspots of disturbance using EOS data. UCS-2, this proposal, pp. I-53–55.
- Knuth DE (1973). *The Art of Computer Programming: Volume 1, Fundamental Algorithms*, Second Edition. Addison-Wesley, Reading, Massachusetts.
- Kulldorff M (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.
- Kulldorff M, Friedlander D, Patil GP, and Taillie C (2002). Syndromic surveillance for biosurveillance and for biosecurity. PCS-13, this proposal, pp. I-63–64.
- Kulldorff M and Nagarwalla N (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14**, 799–810.
- Lehmann EL (1986). *Testing Statistical Hypotheses*, Second Edition. Wiley, New York.
- Mortensen D and Rathbun S (2002). Early detection and delineation of outbreaks of invasive plant species. UCS-2, this proposal, pp. I-22–26.
- Myers WL, Kurihara K, Patil GP, and Vraney R (2002). Finding upper level sets in cellular surface data using echelons and SaTScan. Invited Paper, Joint Statistical Meetings on Statistics in an Era of Technological Change, New York City, NY. <http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0801.pdf>
- Myers WL, Kurihara K, Patil GP, and Vraney R (2003). Finding upper level sets in cellular surface data using echelons and SaTScan. *Environmental and Ecological Statistics*. (To appear). <http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0801.pdf>
- Myers WL, Patil GP, and Joly K (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, **4(2)**, 131–152.
- Myers WL, Patil GP, and Taillie C (1999). Conceptualizing pattern analysis of spectral change relative to ecosystem status. *Ecosystem Health*, **5(4)**, 285–293.
- Neggers J and Kim HS (1988). *Basic Posets*. World Scientific, Singapore.
- Patil GP (2002). Next Generation of Potential Outbreak Detection and Prioritization System. Invited comment and discussion, National Syndromic Surveillance Conference, New York

City, September 2002.

<http://www.stat.psu.edu/~gpp/PDFfiles/SyndromicSurveillance%20Comment.pdf>

Patil GP (2003). Invited Organizer, Chair and Speaker for a Special Invited Session on Hotspot Detection, Delineation, and Prioritization for Geographic Surveillance and Early Warning Systems, Joint Statistical Meetings on Bridge to Discovery and Knowledge, San Francisco, CA.

Patil GP, Balbus J, Biging G, JaJa J, Myers WL, and Taillie C (2003). Multiscale Advanced Raster Map Analysis System: Definition, Design and Development. *Environmental and Ecological Statistics*. (to appear). <http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0203.pdf>

Patil GP, Bishop J, Myers WL, Taillie C, Vraney R., and Wardrop DH (2002). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. Invited Paper, International Society for Ecosystem Health, Washington, DC. <http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0501.pdf>

Patil GP, Bishop J, Myers WL, Taillie C, Vraney R, and Wardrop DH (2003). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. *Environmental and Ecological Statistics* (to appear). <http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0501.pdf>

Patil GP, Brooks RP, Myers WL, Rapport DJ, and Taillie C (2001). Ecosystem health and its measurement at landscape scale: Towards the next generation of quantitative assessments. *Ecosystem Health*, **7(4)**, 307–316.

Patil GP, Brooks RP, Myers WL, and Taillie C (2002). Multiscale advanced raster map analysis system for measuring ecosystem health at landscape scale—A novel synergistic consortium initiative. In *Managing for Healthy Ecosystems*, D. Rapport et al. (eds), Lewis Publishers, Washington DC. pp. 567–576.

Patil GP, Myers WL, Taillie C, and Wardrop D (2002). Hotspot Detection and Early Warning for Synoptic and Network-Based Syndromic Surveillance. Invited Poster Presentation, National Syndromic Surveillance Conference, New York City, September 2002. <http://www.stat.psu.edu/~gpp/PDFfiles/Poster%201.pdf>

Patil GP and Taillie C (2002). Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking methods. Invited Paper, International Environmetrics Society, Genova, Italy. <http://www.stat.psu.edu/~gpp/PDFfiles/TR2001-1204.pdf>

Patil GP and Taillie C (2003). Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking methods. *Environmental and Ecological Statistics*, 2003(to appear). <http://www.stat.psu.edu/~gpp/PDFfiles/TR2001-1204.pdf>

Phoha S, Patil GP, Friedlander, D., Taillie C, and Yen J (2002). Tasking of a self-organizing surveillance mobile sensor networks. UCS-4, this proposal, pp. I-73–77.

Press WH, Teukolsky SA, Vetterling WT, and Flannery BP (1992). *Numerical Recipes in C*, Second Edition. Cambridge University Press, Cambridge.

Ray A and Phoha S (2002). A language measure for discrete-event automata. Proc. of the International Federation of Automatic Control (IFAC) World Congress b'02, Barcelona, Spain, July 2002.

- Salem F and Kafatos M (2001). Hyperspectral image analysis for oil spill mitigation. 22nd Asian Conference Remote Sensing, November 5–9, 2001, Singapore.
- Sarangan V and Acharya R (2001). A study on using network flows in hierarchical QoS routing. *Globecom 2001*, San Antonio, November 2001.
- Sarangan V, Ghosh D, and Acharya R (2002). State aggregation using network flows for stochastic networks. *Globecom 2002*, Taipei, November 2002.
- Shalizi CR, Shalizi KL, and Crutchfield JP (2002a). Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. *Journal of Machine Learning Research* (submitted). Santa Fe Institute Working Paper 02-10-060.
- Shalizi CR, Shalizi KL, and Crutchfield JP (2002b). Pattern discovery in time series, Part II: Implementation, evaluation, and comparison. *Journal of Machine Learning Research* (submitted). Santa Fe Institute Working Paper 02-10-XXX.
- Trotter WT (1992). *Combinatorics and Partially Ordered Sets*. Johns Hopkins University Press, Baltimore.
- Waller L (2002). Methods for detecting disease clustering in time or space. In *Statistical Methods and Principles in Public Health Surveillance*. R. Brookmeyer and D. Stroup, eds. Oxford University Press (to appear).
- Wang X and Ray A (2002). Signed real measure of regular languages. Proc. of the American Control Conference, Anchorage, Alaska, May 2002.
- Wardrop D, Myers WL, Patil GP, and Taillie C (2002). Network-based analysis of biological integrity in freshwater streams. UCS-1, this proposal, pp. I-4–10.
- Yang K-S, Yang R, and Kafatos M (2001). A feasible method to find areas with constraints using hierarchical depth-first clustering. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, L. Kerschberg and M. Kafatos (eds). pp. 257–262.
- Yang R, Yang K-S, Kafatos M, and Wang XS (2001). Value range queries on earth science data via histogram clustering. In *Interim Proceedings of the International Workshop in Temporal Data Mining, TSDM20000*, Lyon, France. *Lecture Notes in Artificial intelligence*, Springer, New York.