

**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**

**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**

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 ε -complexity and ε -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.

1. Introduction and Motivation

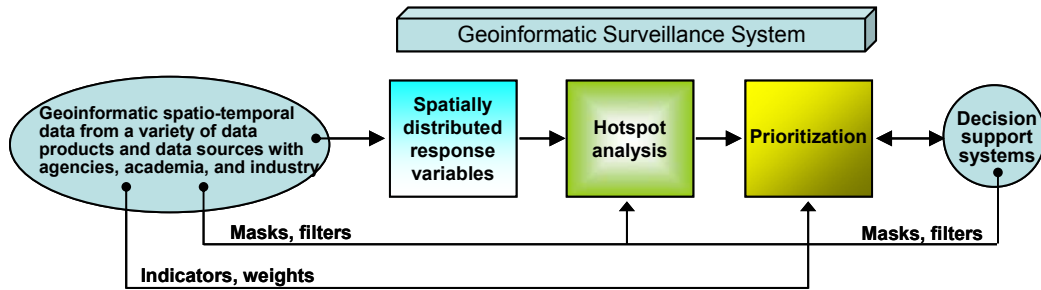
A primary purpose of this proposal is to invent, implement, and interface innovative information technology (IT) for the much needed geoinformatic surveillance decision support system for hotspot detection and prioritization in the project-networked virtual situation room capable of online interaction, 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.

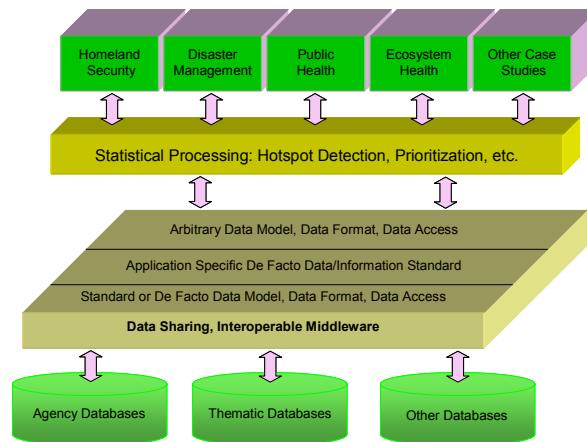
Clearly, clusters are clusters. They can be of any shape, and cannot be captured only by circles. This is likely to give more false alarms and more false negatives than warranted. What we need is the capability to detect arbitrarily shaped clusters and the ability to handle network-based as well as cell-based 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).



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 ε -complexity and ε -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.

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.

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 ε -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 research¹ 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).

¹ 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.

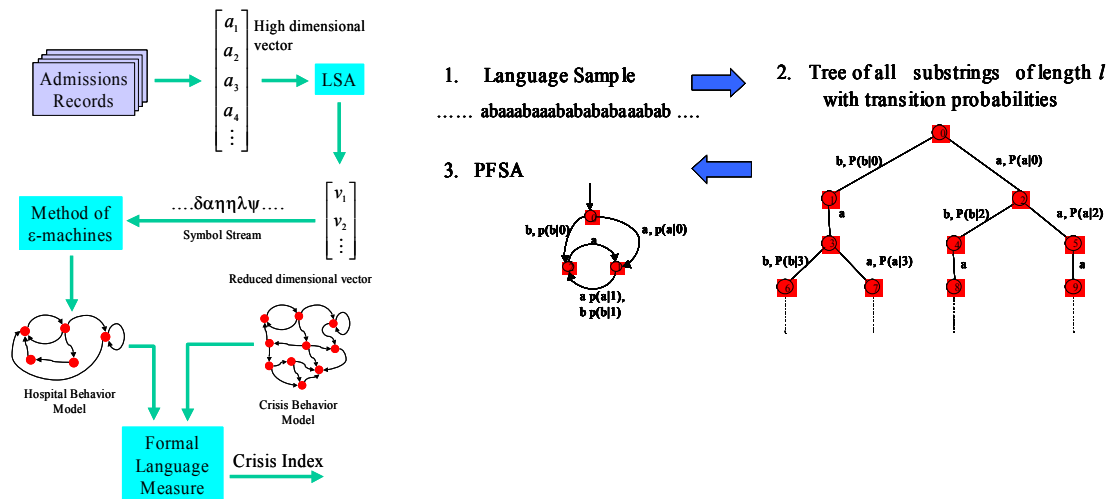


Figure 1. (left) The overall procedure, leading from admissions records to the crisis index for a hospital. The hotspot detection algorithm is then applied to the crisis index values defined over the hospital network. **(right)** The ϵ -machine procedure for converting an event stream into a parse tree 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 ϵ -machines models the behavior associated with an event stream by creating a probabilistic finite state machine that can recognize the substrings in 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.

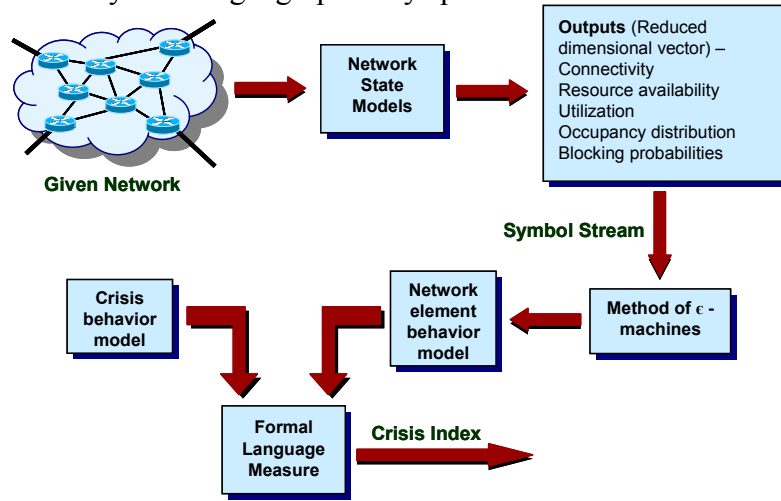


Figure 2. Procedure for obtaining the crisis index for network elements.

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