J. Ind. Soc. Agril. Statist.
61(2), 2007: 132-146

Statistical Geoinformatics of Geographic Hotspot Detection and Multicriteria Prioritization for Monitoring, Etiology, Early Warning, and Sustainable Management for Digital Governance in Agriculture, Environment, Ecology and Ecohealth

Ganapati P. Patil
Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Penn State University, University Park, PA 16802, USA

SUMMARY

This paper is based on the Inaugural Session Keynote Address to the Conference, given in the spirit of inviting the attention of the audience to some of the initiatives of the author that have presently culminated into a novel and innovative project for digital governance and hotspot geoinformatics under the sponsorship of the US National Science Foundation.

GeoInformatics of geospatial and spatio-temporal hotspot detection and prioritization is a critical need for the 21st Century. A declared need is around for statistical geoinformatics and software infrastructure development. A hotspot can mean an unusual phenomenon, anomaly, aberration, outbreak, elevated cluster, critical area. The declared need may be for monitoring, etiology, early warning, or sustainable management. The responsible factors may be natural, accidental or intentional. The five year NSF Digital Government Research Program project has been instrumental to conceptualize hotspot geoinformatics partnership among several interested cross-disciplinary scientists in academia, agencies, and communities around the world. Our efforts are driven by a wide variety of case studies involving a wide variety of critical societal issues.

You are invited to participate in ongoing workshop series around the world in a manner most productive for your purposes and publications. You will have the opportunity to strengthen, advance, and accelerate your in-house research workplan involving novel geoinformatics and innovative hotspot dynamics with capability for early warning and sustainable management. It will be a pleasure to communicate, interact and publish. See the website :

http://www.stat.psu.edu/hospots/pdfs/OverallInfo_ShortCourseandWorkshops.pdf

Key words: Hotspot detection and prioritization, Surveillance system, Digital government, GIS, Information technology, Upper level set scan statistic, Hasse diagrams, Partially ordered sets, Hotspot rating, Carbon budgets, Water resources, Ecosystem health, Public health, Drinking water distribution system, Persistent poverty, Environmental justice, Crop pathogens, Invasive species, Biosurveillance, Remote mobile sensor network, Early warning system.

1. SETTING THE STAGE

It is a great pleasure for me to be here at this historic Diamond Jubilee Program Conference, also in honor of Dr. V.G. Panse and Dr. P.V. Sukhatme. I was fortunate to have their friendship and mentorship for many years starting from 1954 when I was a 20 year old graduate student lucky enough to have a journalistic pass for 1954 Baroda Congress, journalistically accessing everyone that looked important! It was wonderful to get to know these two wonderful people, solid professionals and solid humanbeings, caring and affectionate.

It is also a great pleasure for me to be here following in the footsteps of a living statistical legend such as Dr. C.R. Rao. His talk has emphasized data mining, and

I trust that everyone around has now the data mining mind set. It is not that we have not been doing data mining, but perhaps not on the scale of what is understood to be data mining today.

And lastly, it is a profound pleasure for me to be introducing you to the exciting initiative of digital governance and hotspot geoinformatics and related developments! This has been a gold mine indeed, also in the context of knowledge society and knowledge economy in which we find ourselves today.

To begin with, consider the following three stimulating scenarios followed by a brief overview of the initiative in digital governance and hotspot geoinformatics.

(a) Statistics and Significance

Science strives for the discovery of significant Scientific Truth. It is Statistics that takes care of the uncertainty of the Scientific Method consisting of design, analysis, and interpretation, and even the assessment of significance. The society in which we live has chosen to fully use Statistics as a decisive instrument to deal with societal crises, whether they be related to environment, education, economy, energy, engineering or excellence. While it is exciting that we are alive in the age of information, and while it is unfortunate that we find ourselves in the crisis of environment, it is only a bliss to have the opportunity to more effectively serve the cross-disciplinary cause of statistics, ecology, environment, and society in the research, training, and outreach setting.

(b) Raster Map and Change Map

What message does a remote sensing-derived land cover land use map have about the large landscape it represents? And at what scale and at what level of detail?...Does the spatial pattern of the map reveal any societal, ecological, environmental condition of the landscape? And therefore can it be an indicator of change?...How do you automate the assessment of the spatial structure and behavior of change to discover critical areas, hotspots, and their corridors?...Is the map accurate? How accurate is it? How do you assess the accuracy of the map? Of the change map over time for change detection? What are the implications of the kind and amount of change and accuracy on what matters, whether climate change, carbon emission, water resources, urban sprawl, biodiversity, indicator species, or early warning? And with what confidence, even with a single map/change-map? ...Research is expected to find answers to these questions and a few more that involve multicategorical raster maps based on remote sensing and other geospatial data. It is also expected to design a prototype advanced raster map analysis system for digital governance.

(c) Surveillance GeoInformatics and Digital Governance

Geoinformatic surveillance for spatial and temporal hotspot detection and prioritization is a critical need for the 21st century Digital Government. A hotspot can mean an unusual phenomenon, anomaly, aberration, outbreak, elevated cluster, or critical area. The declared need may be for monitoring, etiology, management, or early warning. The responsible factors may be natural, accidental or intentional, with relevance to both infrastructure and homeland security. This involves critical societal issues, such as carbon budgets, water resources, ecosystem health, public health, drinking water distribution system, persistent poverty, environmental justice, crop pathogens, invasive species, biosecurity, biosurveillance, remote sensor networks, early warning and homeland security. The geosurveillance provides an excellent opportunity, challenge, and vehicle for synergistic collaboration of computational, technical, and social scientists.

(d) Brief Overview of the Initiative of Digital Governance and Hotspot GeoInformatics

This initiative describes a multi-disciplinary research program based on novel methods and tools for hotspot detection and prioritization, driven by a wide variety of case studies of direct interest to several government agencies. These case studies deal with critical societal issues.

Our methodology involves an innovation of the popular circle-based spatial scan statistic methodology. In particular, it employs the notion of an upper level set and is accordingly called the upper level set scan statistic, pointing to the next generation of a sophisticated analytical and computational system, effective for the detection of arbitrarily shaped hotspots along spatiotemporal dimensions. We also propose a novel prioritization scheme based on multiple indicator and stakeholder criteria without having to integrate indicators into an index, using revealing Hasse diagrams and partially ordered sets.

Responding to the Government's role and need, we propose a cross-disciplinary collaboration among federal agencies and academic researchers to design and build the prototype system for surveillance infrastructure of hotspot detection and prioritization. The methodological toolbox and the software toolkit developed will support and leverage core missions of federal agencies as well as their interactive counterparts in the society. The research advances in the allied sciences and technologies necessary to make such a system work are the thrust of this initiative. A multi-disciplinary multi-institution research team will address the issues in an integrated manner, a crucial element of success. The team comprises several leading researchers with track records from research universities. Information technologies promise to make Government more efficient and responsive. The purpose of this initiative is to help that happen.

2. MOTIVATION, INTRODUCTION AND JUSTIFICATION

We propose a multi-disciplinary research program to develop infrastructure for geoinformatic surveillance based on novel methods and tools, tightly coupled with case studies of critical importance to several government agencies. In particular, we propose to enhance and broaden the popular spatial scan statistic method which has been widely used for medical surveillance. For example, during the summer of 2001, it was successfully used for the early detection of dead bird clusters to localize West Nile virus epicenters in New York City. Cluster findings led to preventive measures such as targeted application of mosquito larvicide (Mostashari *et al.* 2003). Our enhancement is called the upper level set (VLS) scan statistic (Patil 2002; Patil *et al.* 2004; Myers *et al.* 2006; Patil *et al.* 2004; Patil *et al.* 2004; Patile and Taille 2004a). Some of its attractive features include :

1. identification of arbitrarily shaped clusters
2. data-adaptive zoning of candidate hotspots
3. applicable to data on a network
4. yields both a point estimate and a confidence set for the hotspot
5. uses hotspot-membership rating to map hotspot boundary uncertainty
6. computationally efficient
7. applicable to both discrete and continuous syndromic responses
8. identifies arbitrarily shaped clusters in the spatial-temporal domain and
9. provides a typology of space-time hotspots with discriminatory surveillance potential

The ULS scan statistic ranks hotspots according to 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, political and social considerations, etc. We therefore envision a suite of indicator values attached to each hotspot with large indicator values signifying greater importance. Different indicators reflect different criteria and may rank the hotspots differently. Therefore, we also propose a prioritization tool based on multiple indicator and stakeholder criteria without having to subjectively integrate indicators into an index. The prioritization tool employs Hasse diagrams for visualization purposes and partially ordered set for analytical purposes (Patil and Taillie 2004b).

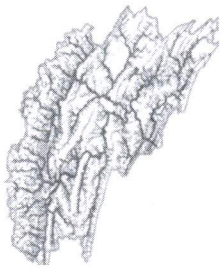
Our team involves researchers with a solid track record in a number of complementary areas that are at the core of this project. Our approach will develop and combine appropriate methodologies paying particular attention to the related computational aspects. We will integrate the resulting advances into a decision support system to be used on a rich set of large-scale case studies. The project goals and results will be achieved in a well-integrated disciplinary and cross-disciplinary effort coupled with matching educational abilities.

3. ILLUSTRATIVE APPLICATIONS AND CASE STUDIES

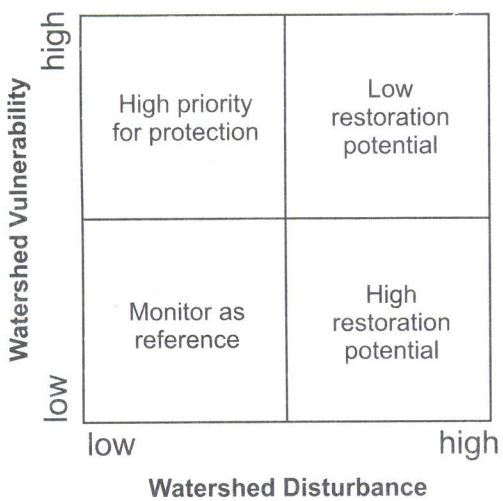
The proposed geosurveillance project identifies studies in health, environment, persistent poverty, environmental justice on the one hand, and in biosurveillance, crop surveillance, and security on the other. This section describes these illustrative applications and case studies.

Network analysis of biological integrity in freshwater streams: This study will employ the network version of the upper level set scan statistic to characterize biological impairment along the rivers and streams of Pennsylvania and to identify subnetworks that are badly

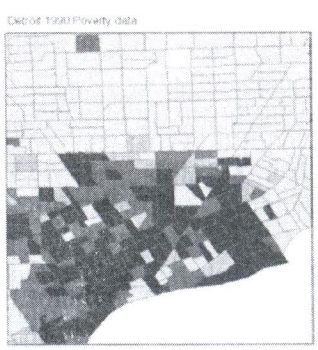
impaired. The state Department of Environmental Protection is determining indices of biological integrity (IBI) at about 15,000 sampling locations across the Commonwealth. Impairment will be measured by a complemented form of these IEI values. We will also use remotely sensed landscape variables and physical characteristics of the streams as explanatory variables in an attempt to account for impairment hotspots. Hotspots that remain unaccounted for after this filtering exercise become candidates for more detailed modeling and site investigation.



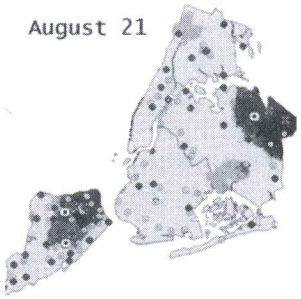
Watershed prioritization for impairment and vulnerability: This study will develop a prioritization model for watersheds (12-digit HUCs) of the Mid-Atlantic Highlands. A suite of indicators will be identified to assess each watershed's susceptibility to impairment (vulnerability). A second suite of indicators will measure actual stress or disturbance for each watershed. The watersheds will then be ranked according to each of the two separate sets of indicators. The proposed prioritization methodology will be used for ranking purposes. Each watershed is thus assigned a pair of ranks indicating its vulnerability status and its disturbance status. The pairs of ranks yield a scatter plot in the disturbance × vulnerability plane. The four quadrants in this plot have distinctly different management implications, as depicted in the accompanying diagram. Disturbance will be measured by stressor variables such as excess sediment, riparian degradation, mine drainage, excess nutrients, exotic species, agriculture (esp. on slopes), road crossings, forest fragmentation, and indices biological impairment. Vulnerability primarily reflects physical characteristics and natural features of the watershed and can be measured by hydrogeomorphology (HGM), climate, aspect, slope, stream sinuosity, soil type, bedrock, and water source. Products include a procedure for classifying watersheds by their features and condition, a taxonomy of MidAtlantic watersheds, and a set of monitoring and restoration options for each watershed class that can assist managers in developing TMDL (total maximum daily load) plans.



Spatial-temporal patterns of poverty in US metropolitan areas: Poverty has been a persistent problem for the US and a costly target of federal policy interventions for many decades. This study is driven by four questions concerning urban poverty: (1) What explains the persistence of poverty, over time? (2) What explains the growth of high poverty neighborhoods? (3) What explains the geographic concentration of the poor? (4) How have policy interventions affected the patterns of urban poverty? We hypothesize that the explanations of urban poverty will vary, depending on the different patterns of persistence, growth and concentration, and that examination of these patterns will provide clues for improved policy interventions. A principal information source will be the 1970-2000 census tract data with boundaries rectified for temporal comparisons. Approximately 45,000 metropolitan tracts have complete poverty data for all four census years. We will employ the proposed ULS scan statistic to identify Y space-time clusters of metropolitan poverty, to track their time-slice trajectories, and to develop a spatial-temporal typology for metropolitan poverty in the US. Poverty is a household, instead of a per capita, characteristic so appropriate modifications will be made to the scan statistic methodology to account for statistical clustering and variable household sizes.

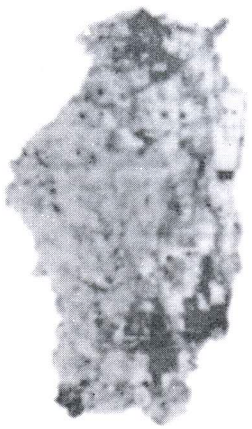


Dead bird clustering - early warning system for West Nile virus : Since the 1999 West Nile (WN) virus outbreak in New York City (NYC), health officials have been searching for an inexpensive and real-time early warning system that could signal increased risk of human

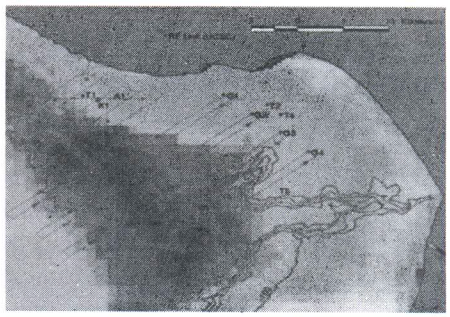


WN infection, and provide a basis for targeted public education and increased mosquito control. Laboratory evidence of WN virus preceded most human infections in 2000 but sample collection and laboratory testing are time-consuming and costly. We have evaluated the cylinder-based space-time scan statistic for detecting small area clustering of dead bird reports and have found it useful in providing an early warning of West Nile virus activity in NYC. All unique non-pigeon dead bird reports were geocoded, and categorized as "cases" if occurring in the prior 7 days, "controls" if occurring during a historic baseline, or censored. The proposed case study would revisit the analysis using the ULS space-time scan statistic. Since the latter allows for arbitrarily shaped clusters in both the spatial and temporal dimensions, there is potential for earlier detection with more accurate delineation as well as a reduced false alarm rate.

Mapping priority hotspots of vegetative disturbance for carbon budgets : Hotspot detection can complement existing approaches to remote measuring and mapping vegetation disturbance for global change research. Existing data products either strive to reduce 'false alarms' by relying on multi-year comparisons of matched 'best quality' data or restrict information to one type of disturbance (e.g., MODIS fire products). National and global carbon budgets, at time scales relevant to inversion of atmospheric transport models, require data that are both more timely and more comprehensive. Producing such data in an operational mode would be well beyond the scope of this case study. Nonetheless it is vital to investigate approaches that could fill this critical gap. The proposed toolkit for hotspot detection and ranking shows great promise for identifying significant disturbance events and providing a 'front-end' to a collaborative system for characterizing their carbon cycle consequences. This case study will sample BOS data streams (primarily from MODIS instruments) and test proposed hotspot algorithms for their value in carbon cycle research and potential for support of carbon management decisions and technology.

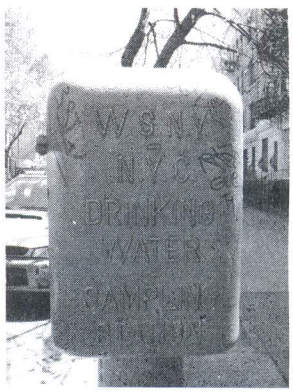


Oceanic surveillance using a remote mobile sensor network: This study will validate empirical methods for dynamic feedback in sensor networks including biological, chemical and physics-based mechanisms. Our application is the mapping of oceanographic fields such as



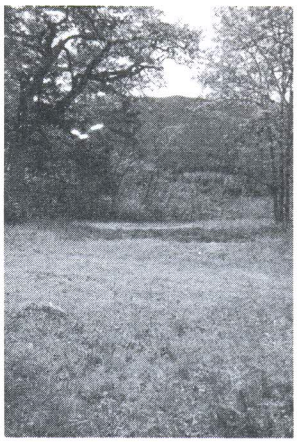
bathymetry, temperature and currents using unmanned undersea vehicles. Upper level set scan statistic theory will be used to guide the vehicles by estimating the location of hotspots based on the data previously taken by the surveillance network. In our case, hotspots are areas of high variation in the data fields. By detecting only the significant variations, resources are not wasted on mapping areas of little change. As mobile sensor platforms move toward estimated hotspot locations, more data will be taken and used to update the locations. The Autonomous Ocean Sampling Network Simulator will be used for high resolution, spatio-temporally coordinated surveys. Oceanographic data fields will be determined by the Harvard Ocean Prediction System.

Surveillance of NYC drinking water distribution system : New York City has installed 892 drinking water sampling stations across the five boroughs. Each 4.5-foot high station is located outdoors and draws water from a nearby water main. The purpose is to monitor general water quality, detect potential health threats, and thwart bioterror activity. Sampling frequency was increased after the 9/11 attacks and, currently, about 47,000 water samples



are analyzed annually. Parameters analyzed include bacteria, chlorine, pH, inorganic and organic pollutants, color, turbidity, odor, and many others. The network version of the ULS scan statistic will provide a real-time surveillance system for detecting and evaluating water quality hotspots within the distribution system.

Early detection of biological invasions. Intentional and unintentional introductions of nonnative exotic species have major economic and ecological impacts across the USA National Academy of Sciences report (2002) estimates the cost of lost crops and containment measures at $137 billion per year. Early detection of invasive weedy plants is the only cost-effective and tractable option for their containment or eradication. But systems for synthesizing on-the-ground observation, spatial data, and newly acquired remotely sensed data are lacking. We propose to apply the ULS scan statistic and prioritization tools to obtain more efficient surveys for invasive species and to improve the responsiveness of environmental managers to outbreaks. Japanese stiltgrass, Microstegium vimineum, has become established in forests and waterways in the eastern US and threatens to significantly reduce forest and riparian species diversity, and impede water flow in rivers and streams. Data are being collected to document the distribution of this species, but often locally established populations have begun to spread before those populations have been detected and likelihood for successful management is severely compromised. Coupling the data resources with the scan statistic represents a promising approach to preventing the transition of invasive plants from isolated established populations to spreading ones, like that depicted in the photograph.



4. DEVELOPMENT OF FUNDAMENTAL METHODOLOGIES AND COMPUTATIONAL TECHNIQUES

4.1 Scan Statistic Methodology

Three central problems arise in geographical surveillance for a spatially distributed response variable. These are (i) identification of areas having exceptionally high (or low) response, (ii) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant, and (iii) assessment of explanatory factors that may account for the elevated response. Although a wide variety of methods have been proposed for modeling and analyzing spatial data (Cressie 1991), the spatial scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) has quickly become a popular method for detection and evaluation of disease clusters. When applied in space-time, the scan statistic can provide early warning of disease outbreaks and can monitor the spatial spread of an outbreak. With innovative modifications, the scan statistic approach can be used for hotspot analysis in any field. We propose to develop methodology and corresponding software for applications of the scan statistic to critical areas of concern for the digital government of the 21st century.

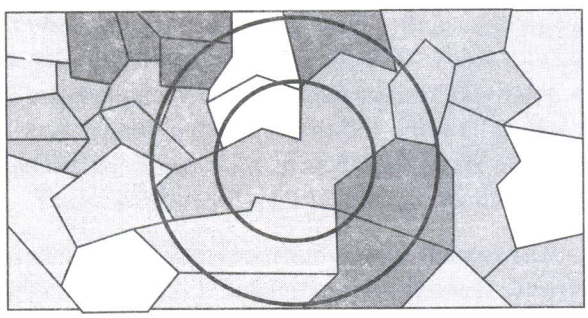
Spatial scan statistic background : The spatial scan statistic deals with the following situation. A region R of Euclidian space is tessellated or subdivided into cells that will be labeled by the symbol a. Data is available in the form of a count Y_a (non-negative integer) on each cell a. In addition, a "size" value Aa is associated with each cell a. The cell sizes Aa are regarded as known and fixed, while the cell counts Y_a are random variables. In the disease setting, the response Y_a is the number of diseased individuals within the cell and the size Aa is the total number of individuals in the cell. Generally, however, the size variable is adjusted for factors such as age, gender, environmental exposures, etc., that might affect incidence of the disease. The disease rate within the cell is the ratio Y_a / Aa. The spatial scan statistic seeks to identify "hotspots" or clusters of cells that have an elevated rate compared with the rest of the region, and to evaluate the statistical significance (p-value) of each identified hotspot. These goals are accomplished by setting up a formal hypothesis-testing model for a hotspot. The null hypothesis asserts that there is no hotspot, i.e., that all cells have (statistically) the same rate. The alternative states that there is a cluster Z such that the rate for cells in Z is higher than for cells outside Z.

An essential point is that the cluster Z is an unknown parameter that has to be estimated. Likelihood methods are employed for both the estimation and significance testing. Candidate clusters for Z are referred to as zones. Ideally, maximization of the likelihood should search across

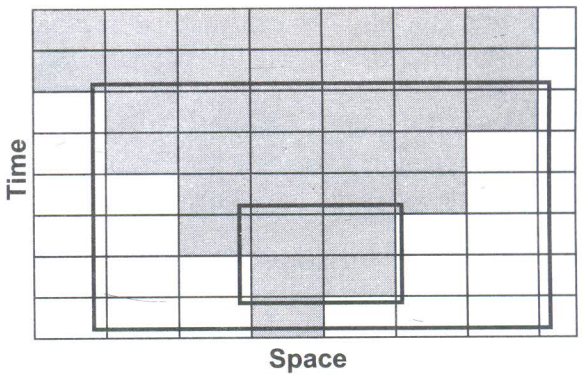
all possible zones, but their number is generally too large for practical implementation. Various devices (e.g., expanding circles) are employed to reduce the list of candidate zones to manageable proportions. Significance testing for the spatial scan statistic employs the likelihood ratio test, however, the standard chi-squared distribution cannot be used as reference or null distribution-in part because the zonal parameter Z is discrete. Accordingly, Monte Carlo simulation (Dwass 1957) is used to determine the needed null distributions.

Explication of a likelihood function requires a distributional model (response distribution) for the response Y_a in cell a. This distribution can vary from cell to cell but in a manner that is regulated by the size variable Aa. Thus, Aa enters into the parametric structure of the response distribution. In disease surveillance, response distributions are generally taken as either binomial or poisson, leading to comparatively simple likelihood functions. The scan statistic that we propose allows continuous response distributions and complex likelihood functions.

Limitations of current scan statistic methodology: Available scan statistic software suffers from several limitations. First, circles have been used for the scanning window, resulting in low power for detection of irregularly shaped clusters (Fig. 1). 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, reflecting the epidemiological origins of the spatial scan statistic, 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 propose to address all these limitations.



Cholera outbreak along a river flood-plain.
Small circles miss much of the outbreak.
Large circles include many unwanted cells.



Outbreak expanding in time
Small cylinders miss much of the outbreak
Large cylinders include many unwanted cells

Fig. 1. Circular spatial scan statistic zonation (left) and cylindrical space-time zonation (right)

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

ULS Scan statistic: 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. We will also compare performance of the ULS scanning tool with that of the traditional spatial scan statistic.

The key element here is enumeration of a searchable list of candidate zones Z. A zone is, first of all, a collection of vertices from the abstract graph. Secondly, those vertices should be connected (Fig. 2) because a geographically scattered collection of vertices would not be a reasonable candidate for a "hotspot." Even with this connectedness limitation, the number of candidate zones is too large for a maximum likelihood search in all but the smallest of graphs. We propose to reduce the list of zones to searchable size in the following way. The response rate at vertex a is $G_a = Y_a / Aa$. These rates determine a function a ® t Ga defined over the vertices in the graph.

This function has only finitely many values (called levels) and each level g determines an upper level set U_g defined by $U_g = \{a : Ga \sim g\}$. Upper level sets do not have to be connected but each upper level set can be decomposed into the disjoint union of connected components. The list of candidate zones Z for the ULS scan statistic consists of all connected components of all upper level sets. This list of candidate zones is denoted by QULS. The zones in QULS 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 QULS does not exceed the number of vertices in the abstract graph (e.g., the number of cells in the tessellation). Finally, QULS becomes a tree under set inclusion, thus facilitating computer representation. This tree is called the ULS-tree (Fig. 3); its nodes are the zones Z E QULS and are therefore collections of vertices from the abstract graph. Leaf nodes are (typically) singleton vertices at which the response rate is a local maximum; the root node consists of all vertices in the abstract graph.

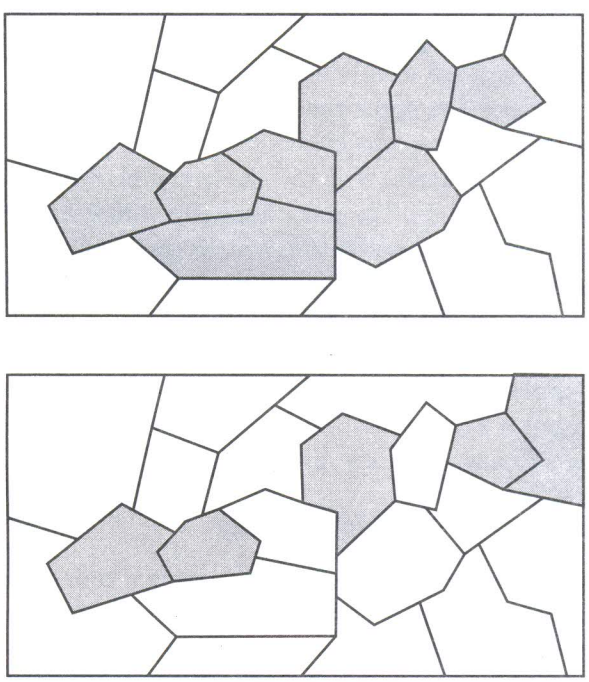


Fig. 2. Connectivity for tessellated regions. The collection of shaded cells on the left is connected and, therefore, constitutes a zone. The collection on the right is not connected.

Finding the collected 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 (Carmen *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: The scan statistic methodology will be extended to include continuous response distributions. 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. Our overall approach is to model the mean and relative variance in terms of the size variable. These moments are functions of the parameters of the response distribution, so that a likelihood function can be written down and parameters estimated by maximum likelihood.

Filtering for explanatory variables: The scan statistic searches for regions of high response relative to a geo-referenced set of prior expected responses. Thus, a hotspot map depicts regions of extreme departure from expectation. The size values Aa which are proportional to model expectations, are the link between the response and potential explanatory variables. In disease surveillance, the Aa are routinely adjusted for factors like age, gender and population size before beginning the analysis (Bithell *et al.* 1995, Kulldorff *et al.* 1997, Rogerson 2001, Waller 2002, Walsh and Fenster 1997, Walsh and DeChello 2001). Such standard, agreed upon, factors are often unavailable in other applications in

which case the initial analysis may identify absolute hotspots by setting all Aa equal to unity. Locations of these highs may provide clues in identifying potential explanatory factors. Next, the size values are adjusted for these factors and the scan statistic is rerun with the adjusted sizes. Comparative configuration of new and old hotspots reveals the impact of these factors on the response under study.

Several methods are available for adjusting the Aa. Suppose, first, that there is only one explanatory variable X. A nonparametric approach partitions the X-values into intervals and calculates the mean response for each interval. The adjusted size value for vertex a is $A\sim = (111_{a}/111)Aa$, where Aa is the old size value, 111a is the mean response for the interval containing a, and 111 is an overall mean response. Regression of Yon X can also be the basis for adjustment provided an appropriate functional relation is identified. Similar approaches work, in principle, for multiple factors. However, the "curse of dimensionality" often comes into play and data sparseness prevents calculation of dependable local means. Our approach, in such cases, is to cluster the data in factor space. A mean response is then calculated for each cluster.

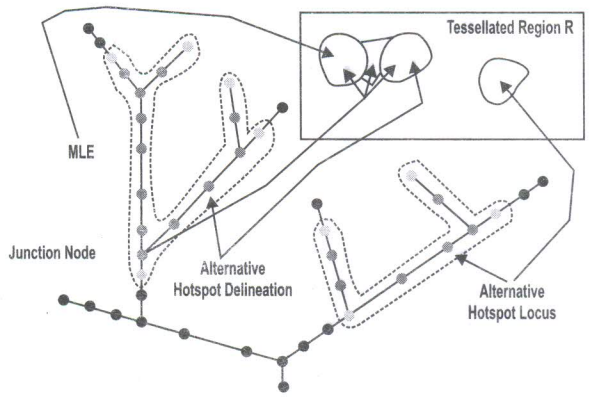


Fig.3. A confidence set of hotspots on the ULS tree. The different connected components correspond to different hotspot loci while the nodes within a connected component correspond to different delineations of that hotspot-**aU** at the appropriate confidence level.

Hotspot confidence sets: The hotspot MLE is 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 (Fig. 3). 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 (Fig. 4). A map of these ratings, with superimposed MLE, provides a visual display of uncertainty in hotspot delineation.

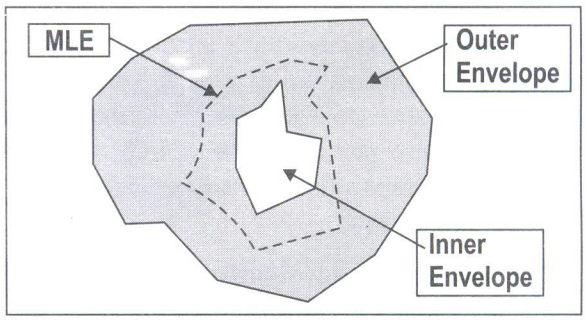


Figure 4. Hotspot-membership rating. Cells in the inner envelope belong to all plausible estimates (at specified confidence level); cells in the outer envelope belong to at least one plausible estimate. The MLE is nested between the two envelopes.

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 statistic employs cylindrical extensions of spatial circles and is unable to detect the temporal evolution of a hotspot (Fig. 1). The space-time generalization of the ULS scan statistic will be able to detect arbitrarily shaped hotspots in space-time. This will allow us to classify space-time hotspots into various evolutionary types-a few of which appear on the left hand side of Fig. 5. The merging hotspot is particularly interesting because, while it comprises a connected zone in space-time, several of its time slices are spatially disconnected.

4.2 Prioritization Methodology

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. The objects can be represented as a cloud of points in indicator space (Filar and Ross 2001), but the different indicators (coordinate axes) typically convey different comparative messages and there is no unique way to

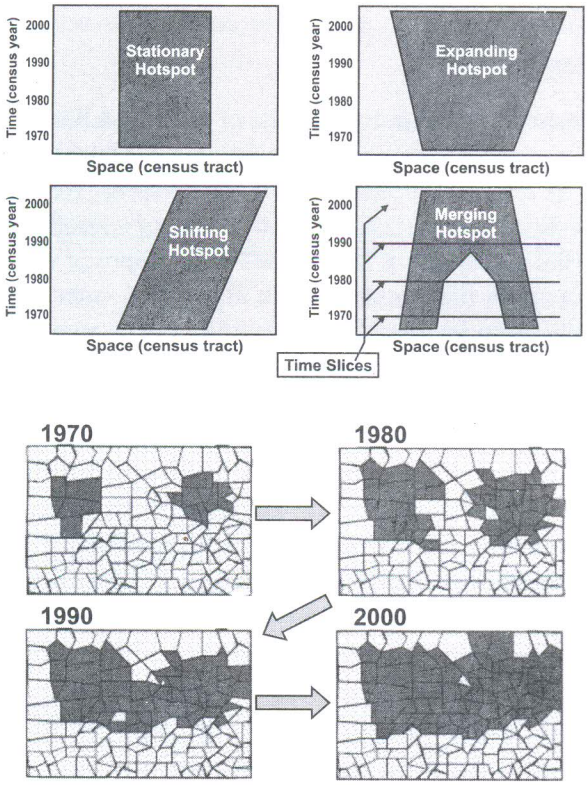


Fig. 5. The four diagrams on the left depict different types of space-time hotspots. The spatial dimension is represented schematically on the horizontal axis while time is on the vertical axis. The diagrams on the right show the trajectory (sequence of time slices) of a merging hotspot.

rank the objects. A conventional solution is to assign 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 or substitutability among indicators. Rather than imposing such a composite, 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 propose to study the collection of all rankings that are compatible with the partial order.

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. Thus, if a and bare hotspots, we say that b is inherently more important than a and we write a -< b if I(a):::; I(b) for all of the indicators 1. If distinct hotspots are distinct in indicator space, the -< relation has the three defining properties of a partial order: (i) transitive: a -< b and b -< c implies a -< c; (ii) antisymmetric: a -< b and b -< a implies a = b; and (iii) reflexive: a -< a. Certain pairs a, b of hotspots may not be comparable under this importance ordering since, for example, there may be indicators such that $1\backslash(a) < 1\backslash(b)$ but $1_2(a) > 1_2(b)$. In this case, hotspot b would be located in the fourth quadrant of Fig. 6. Because of these inherent incomparabilities, there are many different ways of ranking the hotspots while remaining consistent with the importance ordering. A given hotspot a can therefore be assigned different ranks depending upon who does the ranking. It turns out that these different ranks comprise an interval (of integers) called the rank interval of a. Rank intervals can be calculated directly from the partial order. First, define B(a) to be the number of hotspots b for which a -< b, i.e., the count of the first quadrant in Fig. 6. Next, define W(a) as B(a) plus the number of hotspots that are not comparable with a; this is the total count for quadrants 1, 2, and 4 in Fig. 6. The rank interval of a then consists of all integers r such that B(a) £ r £ W(a). The length, W(a) – B(a), of this interval is called the rankambiguity of hotspot a.

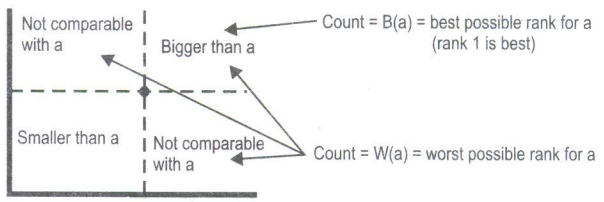
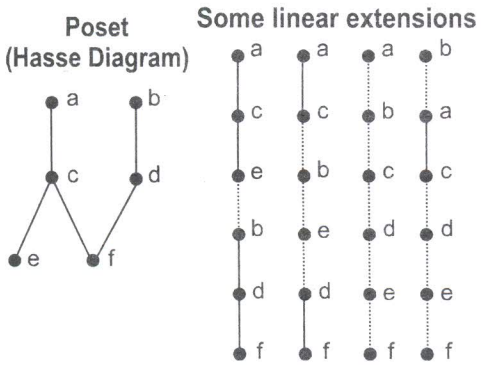


Fig. 6. Regions of comparability and incomparability for the inherent importance ordering of hotspots. Hotspots form a scatterplot in indicator space and each hotspot partitions indicator space into four quadrants.

Hasse diagrams and linear extensions: Posets can be displayed as Hasse diagrams (Fig. 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 > - b ; and

(Hi) 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 (Fig. 7). Enumeration of all possible linear extensions can be accomplished algorithmically as follows. The top element of a linear extension can be anyone of the maximal elements of the Hasse diagram. Select anyone 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 (Fig. 7) and each path through the tree from root node to leaf node determines one linear extension.



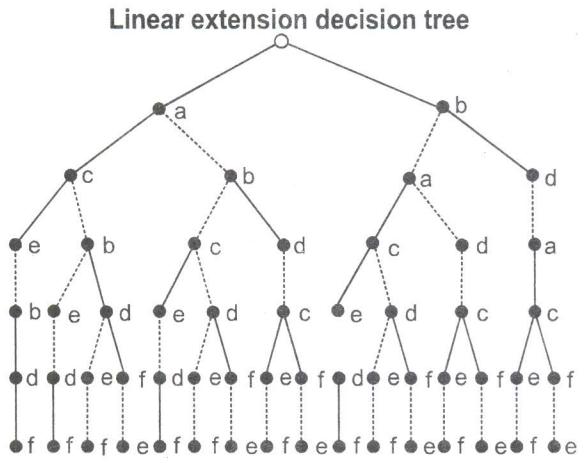


Fig. 7. Hasse diagram of a hypothetical poset, some linear extensions of that poset, and a decision tree enumerating all 16 possible linear extensions. Links shown in dashed (called jumps) are not implied by the partial order. The six members of the poset can be arranged in 6!=720 different ways, but only 16 of these orderings 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 of those hotspots. We ask the question: Is there some objective way of smoothing the partial order into a linear one? Our proposed solution treats each linear extension in Fig. 7 as a voter and we apply the principle of majority rule. Focus attention on some member of the poset, say element a, and ask how many of the voters give a to Rank of 1? Rank of 2? Rank of 3? etc. The results are displayed in Fig. 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 Fig. 8).

	Ranks						
Element	1	2	3	4	5	6	Totals
a	9	5	2	0	0	0	16
b	7	5	3	1	0	0	16
c	0	4	6	6	0	0	16
d	0	2	4	6	4	0	16
e	0	0	1	3	6	6	16
f	0	0	0	0	6	10	16
Totals	16	16	16	16	16	16	

Fig. 8. (Left) Rank-frequency table for the poset of Fig. 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 Fig. 7. The curves are stacked one above the other giving a linear ordering of the elements: a> b > c > d > e > f

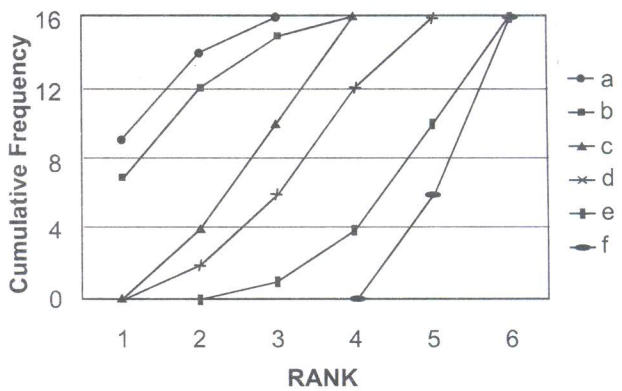


Fig. 9. (Left) Two iterations of the CRF operator are required to transform this partial order into a linear order. (Right) A poset for which the CRF operator produces ties.

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 (Fig. 9). The CRF operator can also produce ties in the final linear ordering.

Prioritization Research Issues: Proposed research 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 because their number is too large. For instance, a recent UNEP HEI poset has only 141 members but the number of linear extensions exceeds 8×10^{105}. As an alternative to 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 Ω of all linear extensions of a given poset. If $O_J \in \Omega$ is the current linear extension, the MCMC transition to the next (proposed) linear extension is accomplished by randomly selecting a jump (see Fig. 7) from O_J and interchanging its two endpoints. We will develop efficient coding to implement this algorithm as well as establish stopping rules and standard errors for the estimated rank-frequency distributions. See Aldous (1987), Brightwell and Winkler (1991) and Haggstrom (2002) for fmiher elaboration of MCMC methods applied to discrete data structures.

- **Ties:** When several hotspots have identical indicator values, they coincide in indicator space and are said to be tied. In case of ties, anti-symmetry fails and the importance ordering is a pre-order instead of a partial order. Our solution is to represent tied hotspots by a single node in the Hasse diagram but, to that node, we also attach the integer count (ramification index) of the number of hotspots represented by the node. Ramification index values must be taken into account in doing the MCMC sampling and in compiling the rank-frequency table. Drawing an analogy with football rankings, if two teams are tied for number one then they collectively consume two ranks and the next team receives rank 3. Note that the CRF operator can produce ties even if there are no ties according to the original suite of indicators.

- **Measurement and Estimation Error:** To put this issue into perspective, suppose there are two indicators and we wish to compare hotspots a and b. Also, suppose $I_1(a) = 10$ and $I_1(b) = 18$ while $I_2(a) = 5$ and $I_2(b) = 4.99$. Strict application of the importance ordering says that the hotspots are not comparable. Nonetheless, their I_2-values are so close (possibly differing only by measurement error) that one might be inclined to order the hotspots according to just I_1. A similar issue arises in applying the MCMC version of the CRF operator since the rank-frequencies must be estimated and are therefore subject to estimation error. In making comparisons, should raw estimates be used or should one use only statistically significant differences? At this stage, we have no prescription for settling these issues but we will explore the multiple comparison and fuzzy comparison literature to develop appropriate procedures. Stochastic frontier analysis (Chames *et al.* 1994, Kumbhakar and Knox 2000) and differential weighting may also lead to a solution.

- **Non-uniform (Weighed) 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. For example, if a particular indicator is especially important, we might scan a linear extension, count the number of links that are consistent with the indicator, and weight in proportion to that count.

5. GEOINFORMATIC SURVEILLANCE 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 effective visualization techniques for hotspot detection and prioritization using statistical methodologies developed in the project. In fact, we have

recently addressed the problem of quickly identifying regions for large scale multivariate maps for which a number of geospatial parameters satisfy certain conditions. See JaJa and Shi (2001). We will extend these techniques in a number of directions suggested by the proposed scanning techniques and prioritization tools.

Information Visualization, User Interface Design, and GIS Linkage

A major goal is to develop a visualization interface integrated with the statistical software tools developed in this project. Infonnation visualization and interface design are critical for effective use of these tools. A phased implementation will allow us to implement simple algorithms at first and then embed more sophisticated algorithms. As our implementations mature, we will conduct usability tests in coordination with the specialists to refine the interfaces and demonstrate efficacy.

6. INTEGRATION OF RESEARCH, EDUCATION AND DISSEMINATION

An essential part of this project is to introduce methods and tools at the core of the upper level scan statistic system to hotspot analysis researchers in various agencies. Constant interactions among the participating researchers and partners will ensure the development of techniques and tools tailored to address the needs of the involved federal agencies and other partners.

In graduate education, we will integrate the techniques and methods into the wide range of related graduate courses offered on the three participating campuses. Graduate students will test and validate various tools as they become available through the project. Also, the graduate students supported on the 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 plan: two monographs, two case books, two thematic journal issues, six-monthly research workshops and tutorials, and distributed information management. These products and outcomes will help evaluate the success of the educational activities of the project. Four REU additions to the project will be requested for interactive undergraduate dimension. The investigators have a strong commitment to the principle of diversity and we will enhance the current collaboration between UMD and Bowie State.

7. TWO CURRENT GEOINFORMATICS MONOGRAPHS

The following two monographs have recently appeared. They deal with statistical GeoInformatics and Geospatial Data Mining.

Monograph 1: Landscape Pattern Analysis for Assessing Ecosystem Condition, Johnson and Patil (2007)

Synopsis: One of our greatest current challenges is the preservation and remediation of ecosystem integrity. This requires monitoring and assessment over large geographic areas, repeatedly over time, and cannot be practically fulfilled by field measurements alone. Remotely sensed imagery plays a crucial role by its ability to monitor large spatially continuous areas. This technology increasingly provides extensive spatial-temporal data; however, the challenge is to extract meaningful environmental information from such extensive data. This book presents a new method for assessing spatial pattern in raster land covering maps based on satellite imagery in a way that incorporates multiple pixel resolutions. This is combined with more conventional single-resolution measurements of spatial pattern and simple non-spatial land cover proportions to assess predictability of both surface water quality and ecological integrity within watersheds of the state of Pennsylvania (USA).

Monograph 2: Pattern-Based Compression of Multi-Band Image Data for Landscape Analysis, Myers and Patil (2007)

This book describes an integrated approach to using remotely sensed data in conjunction with geographic information systems for landscape analysis. Remotely sensed data are compressed into an analytical image-map that is compatible with the most popular geographic information systems as well as freeware viewers. The approach is most effective for landscapes that exhibit a pronounced mosaic pattern of land cover. The image maps are much more compact than the original remotely sensed data, which enhances utility on the internet. As value-added products, distribution of image-maps is not affected by copyrights on original multi-band image data.

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

REFERENCES

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.

Bithell, I.F., Dutton, S.I., Neary, N.M., and Vincent, T. I. (1995). Controlling for socioeconomic confounding using regression methods. *Community Health*, **49**, S15-S19.

Brightwell, G. and Winkler, P. (1991). Counting linear extensions. *Order*, **8**, 225-242.

Charnes, A., Cooper, A.Y., Lewin, A.Y., and Seiford L.M. (1994). *Data Envelopment Analysis: Theory, Methodology, and Applications.* Kluwer, Boston.

Cormen, T.H., Leierson, C.E., Rivest, R.L., Stein, C. (2001). *Introduction to Algorithms.* Second Edition, Cambridge, Massachusetts.

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

Di Battista, G., Eades, P., Tamassia, R. and Tollis, I.G. (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. *Ann. Math. Statist.*, **28**, 181-187.

Filar, I. A. and Ross, N.P. (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.

Fishburn, P.C. (1985). *Interval Orders and Interval Graphs: A Study of Partially Ordered Sets.* Wiley, New Yark.

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.*

Knuth, D.E. (1973). *The Art of Computer Programming Fundamental Algorithms,* Second Edition. Addison-Wesley, Reading, Massachusetts.

Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. - Thoery Methods,* **26**, 1481-1496.

Kulldorff, M., Feuer, E.J., Miller, B.A. and Freedman, L.S. (1997). Breast cancer clusters in Northeast United States: A geographic analysis. *Amer. J. Epidemiology*, **146**, 161-170.

Kulldorff, M. and Nagarwallas, N. (1995). Spatial disease clusters: Detection and inference. *Stat. Med.*, **14**, 799-810.

Kumbhakar, S.C. and Knox Lovell, C.A. (2000). *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* Second Edition, Wiley, New York.

Mostashari, F., Kulldorff, M., Hartman J.J., Miller, J.R., Kulasekera V. (2003). (for the New York City West Nile Virus Surveillance Working Group). Dead bird clusters as an early warning system for West Nile virus activity. *Emerg Infect Dis.*, **9**, 641-646.

Myers, W.L., Kurihara, K., Patil, G.P., and Vraney, R. (2006). Finding upper level sets in cellular surface data using echelons and SaTScan. *Envir. Ecol. Stat.*, **13(4)**, 379-390.

National Academy of Sciences (2002). Predicting invasions of non-indigenous plants and plant pests. *National Research Council*, 198.

Neggers, J. and Kim, H.S. (1998). *Basic Posets.* World Scientific, Singapore.

Patil, G.P. (2002). Next Generation of Potential Outbreak Detection and Prioritization System. Invited comment and discussion, National Syndromic Surveillance Conference, New York City. http://www.stat.psu.edu/~gpp/PDFfiles/SyndromicSurveillance%2OComment.pdf

Patil, G.P., Balbus, J., Biging, G., JaJa, J., Myers, W.L. and Taillie, C. (2004). Multiscale advanced raster map analysis system: Definition, design and development. *Envir. Ecol. Stat.*, **11(2)**, 113-138. http://www.stat.psu.edu/~gppIPDFfiles/TR2002-0203.pdf

Patil, G.P., Bishop, J., Myers, W.L., Taillie, C., Vraney, R. and Wardrop, D.H. (2004). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. *Envir. Ecol. Stat.*, **11(2)**, 139-164. http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0501.pdf

Patil, G.P., Myers, W.L., 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. http://www.stat.psu.edu/~gpp/PDFfiles/Poster%201.pdf

Patil, G.P., and Taillie, C. (2004a). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Env. Ecol. Stat.*, **11(2)**, 183-198. http://www.stat.psu.edu/~gpp/PDFfiles/TR2002-0601.pdf

Patil, G. P., and Taillie, C. (2004b). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking methods. *Env. Ecol. Stat.*, **11(2)**, 199-228. http://www.stat.psu.edu/~gpp/PDFfiles/TR2001-1204.pdf

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in C.* Second Edition, Cambridge University Press, Cambridge.

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

Trotter, W.T. (1992). *Combinatorics and Partially Ordered Sets.* Johns Hopkins University Press, Baltimore.

Waller, L. (2002). Methods for detecting disease clustering in time or space. In: *Monitoring the Health of Population: Statistical Methods and Principles in Public Health Surveillance. R. Brookmeyer and D. Stroup (eds.),* Oxford University.

Walsh, S.J. and DeChello, L.M. (2001). Geographical variation in mortality from systemic lupus erythematosus in the United States. *Lupus*, **10**, 637-646.

Walsh, S.I., and Fenster, I.R. (1997). Geographical clustering of mortality from systemic sclerosis in the Southeastern United States 1981-90. *J. Rheumatalogy,* **24(12)**, 2348-2352.