Digital Governance, Hotspot GeoInformatics, and Sensor Networks for Monitoring, Etiology, Early Warning, and Sustainable Management

G.P. Patil (1), V.D. Patil (2) S.P. Pawde (3), Shashi Phoha (4), Vijay Singhal (5), and Raj Zambre (6)

(1) Center for Statistical Ecology and Environmental Statistics, Department of Statistics, The Penn State University, University Park, PA, USA
(2) District Executive Engineer, Department of Irrigation, Jalgaon, MS, India
(3) Watershed Surveillance and Research Institute, JalaSRI, M.J. College, Jalgaon, MS, India
(4) Information Science and Technology Division, ARL, The Penn State University, University Park, PA, USA
(5) District Collectorate, Jalgaon, MS, India
(6) Erallo Technologies Inc., Littleton, MA, USA

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010. 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.


Technical Report Number 2008-1105
TECHNICAL REPORTS AND REPRINTS SERIES
November 2008
Digital Governance, Hotspot GeoInformatics, and Sensor Networks for Monitoring, Etiology, Early Warning, and Sustainable Management *

G.P.Patil (1), V.D. Patil (2) S.P. Pawde (3), Shashi Phoha (4), Vijay Singhal (5), and Raj Zambre(6)

(1) Center for Statistical Ecology and Environmental Statistics, Department of Statistics, The Penn State University, University Park, PA, USA
   Email address : gpp@stat.psu.edu
(2) District Executive Engineer, Department of Irrigation, Jalgaon, MS, India
   Email address: vdpatilmid@rediffmail.com
(3) Watershed Surveillance and Research Institute, JalaSRI, M.J. College, Jalgaon, MS, India
   Email address : spp.jalasri@gmail.com
(4) Information Science and Technology Division, ARL, The Penn State University, University Park, PA, USA
   Email address : sxp26@psu.edu
(5) District Collectorate, Jalgaon, MS, India
   Email address : vijaysinghal@hotmail.com
(6) Erallo Technologies Inc., Littleton, MA, USA
   Email address : zambre@erallo.com

1. INTRODUCTION AND INVITATION
2. PROBLEM IDENTIFICATION AND CONCEPTUAL BACKGROUND
3. REVIEW OF LITERATURE
4. ILLUSTRATIVE STUDY AREAS
5. METHODS AND TOOLS
6. SENSORS ENVISAGED FOR WATER CONSERVATION
7. PLANS AND EXPECTED RESULTS FOR JALGAON DISTRICT
8. FUTURISTIC VISION
   Acknowledgements
   References

1. INTRODUCTION AND INVITATION

This chapter is prepared in the spirit of inviting the attention of the readership to some of the initiatives of the authors that have presently culminated into a novel and innovative project for digital governance and hotspot geoinformatics. 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. The JalaSRI, Watershed Surveillance and Research Institute, Jalgaon, India has been instrumental to initiate several case studies at the district level.

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/hotspots/pdfs/OverallInfo_ShortCourseandWorkshops.pdf and also the twenty-five referenced websites in it.

2. PROBLEM IDENTIFICATION AND CONCEPTUAL BACKGROUND

Today we find ourselves in the knowledge society and knowledge economy. Digital governance and hotspot geoinformatics is an important part of it. 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, hot spots, 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 multigetarchical 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. See Figure 0.

**NSF Digital Government Surveillance Geoinformatics Project, Federal Agency Partnership and National Applications for Digital Governance.**

**Federal Agency Partnership**
- CDC
- DOD
- EPA
- NASA
- NIH
- NOAA
- USFS
- USGS

**National and International Applications**

- Homeland Security
- Syndromic Surveillance
- Social Networks
- Sensor Networks
- Public Health
- Urban Crime
- Water Management
- Environmental Policy
- Poverty Policy
- Invasive Species
- National and International

**Figure 0. Schematic Diagram**

The following two monographs have recently appeared. They deal with Statistical Geoinformatics and Geospatial Data Mining, and can help with more conceptual and methodological background.

**Landscape Pattern Analysis for Assessing Ecosystem Condition**
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). See Johnson and Patil (2007).

**Pattern-Based Compression of Multi-Band Image Data for Landscape Analysis**
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. See Myers and Patil (2007).

3. REVIEW OF LITERATURE

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 (ULS) scan statistic (Patil, 2002; Patil et al. 2004; Myers et al. 2006; Patil, Balbus et al. 2004; Patil, Bishop et al. 2004; Patil and Taillie, 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.

4. ILLUSTRATIVE STUDY AREAS

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 some of 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 IBI 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 x 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.

5. METHODS AND TOOLS

5.1 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 \( A_a \) is associated with each cell a. The cell sizes \( A_a \) are regarded as known and fixed, while the cell counts \( Y_a \) are random variables. In the disease setting, the response \( Y_a \) is the number of diseased individuals within the cell and the size \( A_a \) is the total number of individuals in the cell. Generally, however, the size variable is adjusted for factors such as age, gender, environmental exposures, etc., that might affect incidence of the disease. The disease rate within the cell is the ratio \( Y_a / A_a \). The spatial scan statistic seeks to identify "hotspots" or clusters of cells that have an elevated rate compared with the rest of the region, and to evaluate the statistical significance (p-value) of each identified hotspot. These goals are accomplished by setting up a formal hypothesis-testing model for a hotspot. The null
hypothesis asserts that there is no hotspot, i.e., that all cells have (statistically) the same rate.

The 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 $A_a$. Thus, $A_a$ enters into the parametric structure of the response distribution. In disease surveillance, response distributions are generally taken as either binomial or Poisson, leading to comparatively simple likelihood functions. 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 (Figure 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.
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: 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 (Figure 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 $G_a = Y_a / A_a$.' These rates determine a function $a \rightarrow G_a$ defined over the vertices in the graph. This function has only finitely many values (called levels) and each level $g$ determines an upper level set $U_g$ defined by $U_g = \{ a : G_a \geq g \}$. Upper level sets do not have to be connected but each upper level set can be decomposed into the disjoint union of connected components. The list of candidate zones $Z$ for the ULS scan statistic consists of all connected components of all upper level sets. This list of candidate zones is denoted by $\Omega_{ULS}$.' The zones in $\Omega_{ULS}$ 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 $\Omega_{ULS}$ does not exceed the number of vertices in the abstract graph (e.g., the number of cells in the tessellation). Finally, $\Omega_{ULS}$ becomes a tree under set inclusion, thus facilitating computer representation. This tree is called the ULS-tree (Figure 3); its nodes are the zones $Z \in \Omega_{ULS}$ and are therefore collections of vertices from the abstract graph. Leaf nodes are (typically) singleton vertices at which the response rate is a local maximum; the root node consists of all vertices in the abstract graph.
Figure 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 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 (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).

Figure 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 all 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 (Figure 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 (Figure 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 (Figure 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 Figure 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.

![Diagram of hotspots](image)

Figure 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.

### 5.2 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 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 b are 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 I. If distinct hotspots are distinct in indicator space, the < relation has the three defining properties of a partial order: (i) transitive: a < band b < c implies a < c; (ii) antisymmetric: a < band 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 I1(a) < I1(b) but I2(a) > I2(b). In this case, hotspot b would be located in the fourth quadrant of Figure 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 Figure 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 Figure 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.

**Figure 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 (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 ≠ 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 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 (Figure 7) and each path through the tree from root node to leaf node determines one linear extension.

**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 Figure 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 \(a\) rank of 1? Rank of 2? Rank of 3? 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).

<table>
<thead>
<tr>
<th>Element</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>(b)</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>(c)</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>(d)</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>(e)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>(f)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

**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.

6. SENSORS ENVISAGED FOR WATER CONSERVATION

6.1 Introduction

The availability of a variety of inexpensive micro-sensors with embedded wireless communications have enabled real-time monitoring of natural phenomena that span temporal and spatial scales. This enables in-situ information fusion for comprehension and scientific prediction of spatial-temporal events, which in turn supports scientific decision models that adapt to predicted events. For example, autonomous networks of unmanned undersea vehicles with embedded sensor systems have been designed to formulate high fidelity newcasts and forecasts of the ocean through time-space coordinated sampling to support collaborative undersea mine-hunting missions (Phoha et. al. 2006, Phoha et. al. 1999). The National Ecological Observatory Network (NEON) is another national effort of the US National Science Foundation to create a national observing system for ecological measurements and monitoring to support research (Schimel 2007). In this section we present recent research on sensor networking architectures that enable in-situ scientific decision making with the goal of exploring possible value added enhancements to current plans of water linking by the JalaSRI project in Jalgaon, India. This research will enable the project to establish the appropriate regional infrastructure for utilizing the transformational power of information to support situation aware adaptive control of natural resources, such as optimal water conservation. Other possible uses of such a network are delineated by the NEON project in areas of land use and agriculture, spatial patterns of climate-change that affect eco-hydrology and bio-geo-chemistry, and bio-diversity (Schimel 2007). The important characteristics of the decision-support sensor network architecture are its quality of fusion support, low total cost of ownership, scalability, portability of nodes, and system dependability.

The architectural design of the infrastructure for an adaptive sensor network has generated a lot of research interest and experimentation. The following subsections discuss some of the design
issues for a cost effective, flexible and reconfigurable sensor network. The major new research addressed here is the fusion driven dynamic adaptation of the decision support network. The paper presents innovative analytical models to support regional decision-making. The methodology is extendible and has the potential of influencing the design of a national scale environment monitoring network such as the INDOFLUX (Srinivasan et.al. 2007).

6.2 Cyberinfrastructure Architecture
A sensor network operates on an infrastructure of sensing, computation, and communication, through which it perceives the evolution of physical dynamic processes in its environment. Sensors require physical interaction with the sensed phenomena and are subject to a number of noise factors. Sensor data is therefore highly correlated in the vicinity of a stimulus. To get reliable performance from individually less reliable sensors, time-critical collaborative inference in the vicinity of a stimulus is necessary to circumvent limitations of sensing, communications, power, and equipment faults. We call this dynamic clustering. Characterizing dynamic events in multiple spatial-temporal scales under operational constraints requires the tactful capture of coarse and fine grained system dynamics. A large resource constrained sensor network must, therefore, dynamically switch from coarse to fine grained topologies to support progressively segmented analyses to localize emerging hotspots, correct spatial-temporal misalignments through statistical analyses, and discover distributed higher level associations of emerging patterns in diverse multisource asynchronous data sources.

A typical sensor network should consist of the following components:

- A variety of sensors types including: acoustic doppler, oxygen optode, ecolab analyzer (NO3, Si, PO4, NH4), turbidity sensor, temperature sensor, IR gas analyzer, pH probes, RH probes, quantum sensors, fluorometer, barometer, wind sensor, rain gauge, pyrometer, soil temperature, soil moisture, sonic anemometer, gas analyzer, etc.
- Sensing actuators - Interact with the environment to gather data.
- Local nodes - Log sensor data. Do some signal processing. Store data temporarily.
- Network - Transfer data from field site to data portal.
- Data portal - Store data permanently. Manage and manipulate data.
- User interface - Manipulate data on portal, download for local processing, or use custom tools.

Spatial and temporal sensor sampling rates will vary greatly. Real-time data interactions are necessary. Some local nodes are remote data logging devices that store information for later retrieval. The network physical layer may use long-range 802.11 and/or cell phone connections. The data portal provides a grid computing environment. Data signatures certify the sensor hardware that produced the original data and provide assurance that the data is not tampered with. The exact processing history of all derived data can be verified using cryptographic primitives. Sensors interact with their environment and degrade over time, leading to loss of precision and/or accuracy. With minimal knowledge of degradation modes, it is possible to detect and compensate for calibration problems. Distributed calibration considering a variety of noise models is described in (Brooks 1996).
Several design issues for such dynamic sensor networks have been addressed in (Phoha et al. 2006). These include:

- Sensor deployment, self-organization and localization
- Purposeful mobility and scalability
- Network routing and protocol design
- Power and resource management, and
- Network security.

6.3 Fusion Driven Design Concept

The dynamic adaptation of the sensor network is necessary to support regional decision-making and action-oriented control. The goal is to formulate analytical models by using the non-stationary statistics of the information dynamics of the sensor data to drive in-situ changes in the network design space as depicted in (Fig. 10). This figure illustrates the concept of closed-loop network control that manipulates the network topology \( \tau \), based on feedback information of evolving statistical patterns \( \mathbf{R} \) derived from sensor data sequences \( \{y_k\} \).

![Network Design Space](image)

Fig. 10. Solution Concept

To enhance quality of data fusion and resilience, a distributed sensor network needs to be adaptively reconfigured, where the network topology is updated in real time based on the spatial-temporal information derived from the ensemble of sensor data. We proceed to present methods for in-situ construction of statistical models of sensor information and fusion processes in the next two sections.

A. Multilevel Fusion Model

The Information space of the dynamical system is represented by spatial-temporal statistics of the ensemble of sensor data. In this context, sensor data fusion is posed as a multi-time-scale problem under the following assumptions: (i) quasi-stationarity over the fast-time scale (i.e., stationary over a sufficiently long duration) and (ii) possible non-stationarity caused by small parametric or non-parametric deviations in the system behavior due to accumulating changes in the slow-time scale. We characterize multi-level fusion processes as hierarchical dynamic processes that detect signal patterns in multiple, diverse, and spatially dispersed sensor data streams at four hierarchical levels of abstraction. Symbolization is the first atomic level of fusion.
akin to feature extraction. It captures causal information, communications and computational patterns embedded in the underlying physics. Higher levels of abstraction represent higher levels of fusion.

B. Models of Information Dynamics
The discrete event dynamics of sensor data is modeled as hybrid multilayered interacting probabilistic automata (Phoha et.al. 1999, Phoha et.al 2002). Continuously varying dynamics capture the physical processes at the lowest level of abstraction while discrete event models integrate sensing, computation and communication events in a formal language representation (Phoha et.al. 2002). A formal language measure has been developed for measuring operational deviations from specified behavioral representations (Ray et.al. 2005.) This analytically captures the structural dynamics of the information space at various levels of abstraction and develops a measure for its deviations from normalcy. In its simplest form, the information space is modeled as cellular automata with its cells representing sensors that are on or off and interacting with neighboring cells organized as a lattice (Brooks et.al. 2002). More complex high fidelity models capture the nonlinear interactive and multi scale dynamics of the sensor network at multiple layers of abstraction (Biswas et.al. 2006, Phoha et.al. 2006) to assess coverage, connectivity and coordination. As discussed below, these models enable the analytical formulation and empirical evaluation of a networked decision support system. However, to rigorously address the inverse problem, we incorporate algorithms for in-situ derivation of statistical characteristics of the information space.

C. In-situ Model Construction and Approximation of Information Statistics
We first formulate mathematical techniques for local processing at a sensor node into semantic information with flexible resolution. We represent nodes by traveling wave packets of sensor energy, enabling a semantic interpretation (Friedlander et.al. 2002):

\[
\phi(\vec{x}, t) = \sum_{\vec{v}, \vec{s}, \vec{w}} a_{\vec{v}, \vec{s}, \vec{w}} \psi_{\vec{v}, \vec{s}, \vec{w}}(\vec{x} - \vec{v}t),
\]

(1) where \( \vec{v} \) is the velocity of the wave packet, \( \vec{x} \) is its position at time \( t \), \( \vec{s} \) is its scale and \( \vec{w} \) is its offset. This representation treats space and time in a symmetric manner, preserving translational invariance for forecasting. Because the wavelet transform divides data into different frequency components and analyzes each component with a resolution matched to its scale, we can directly model the dependent properties of sensor data. Hence, we represent the information originating at a single sensor node as a finite set of wavelet coefficients that change due to interactions with the environment or other nodes.
We discretize this dynamical system, both spatially and temporally, through novel symbolization and nonlinear filtering techniques that preserve the statistical characteristics of the sensor data yielding a reduced order representation of the information space (Ray 2004). Multivariate sensor outputs are converted to univariate symbol sequences by partitioning a compact region in the wavelet coefficient space into finitely many discrete blocks. Each block in the partition is labeled as a symbol of a finite alphabet. As the dynamical system trajectory evolves in time, it travels through various blocks generating a symbol sequence. A hidden Markov model is constructed from the symbol sequence as a finite-state automaton (FSA; Fig. 123). The FSA is constructed based on the principle of sliding block codes (Lind et.al. 1995) and the machine states are defined corresponding to the alphabet of symbols. The proposed Symbolic Dynamic Filtering (SDF) (Ray 2004) technique follows the recursion and input/output structure of Rao-Blackwellised Particle Filtering (RBPF) that is a sequential Monte Carlo Markov chain method (Doucet et.al. 2001). However, unlike RBPF that is constructed as a Markov process on a finite-dimensional state space, SDF is constructed on a finite-state automaton with finite memory. While the state variables in RBPF are real-valued Markov processes, the automaton states of SDF are analogous to discrete energy states in Quantum Statistical Mechanics (Pathria 1996).
The next step is order reduction. For each sensor cluster, a local automaton is derived as a shift space of finite type (i.e., having finitely many forbidden blocks). The resulting Perron-Frobenius operator (i.e., the state transition matrix of the FSA) has an invariant algebraic structure with time dependent parameters. This algebraic structure allows order reduction of the automaton by state merging, conceptually analogous to information marginalization in RBPF (Doucet et al. 2001). The state probability distribution, represented as histograms in Fig. 123 is recursively computed as an approximation of the natural invariant density, which is a fixed point of the local Perron-Frobenius operator. For in-situ data fusion and information compression, this model of the Information Space has the major advantage of providing a succinct statistical characterization of the sensor data with the following benefits:

- **Robustness to noise**—the wavelet coefficients not only help represent patterns at different scales, but also severely mitigate the effects of measurement noise and spurious signals. Thus, probability of occurrence of false symbols, which either degrade accuracy of pattern identification or produce false alarms, is significantly reduced.

- **ID Adaptive resolution**—partitioning based on maximization of the Shannon entropy (Rajagopalan, and Ray 2006) makes regions with more information segmented finer. This resolution is associated to sensing density.

- **Capability for early detection of emergent behaviors** with decreased probability of false alarms due to sensitivity to changes in the underlying dynamical system.

- **Compression of multi-sensor information**, into a code book of short packets of statistical pattern yields high throughput, low latency and error-corrected transmission over a wireless communication network.

- **Real-time execution** on COTS platforms.

In contrast to RBPF and other state estimation methods, our technique does not require an explicit model of the system as it extracts the intrinsic in-situ information directly from sensor data. Freedom to choose an arbitrary starting point makes it robust for statistical pattern discovery. This formalism provides the basic analytical framework for capturing effects of change in the network design space on system evolution.
D. Network Reconfiguration
The network design space is reconfigured to adapt to the information space in a manner that preserves the statistical characteristics (predictability) of the ensemble of original sensor data at each level of fusion. In the following steps, we present how we use these concepts to build a probabilistic theory for fusion based decision support by designing flexible sensor networks that capture change in operational environments:

- Network-centric sensor information is organized as a discrete-event dynamic system of interacting probabilistic automata, where sensor nodes may change their internal states through interactions with other nodes or the environment. Sensor nodes generate multivariate asynchronous data streams that interact over the network. Based on these interactions, some sensors may form collaborative clusters.

- The symbolization and filtering processes (fusion levels 0 and 1) for a multivariate stream of asynchronous sensor data are said to be effective to the extent that they preserve the statistics of the original data. The goal here is to design flexible network topologies for sufficiently fine-grained adaptive sensing that can detect changes in the statistics of the information space in emerging hotspots.

- Statistical invariance, simultaneously in space and time, is used to reduce the order of the nonlinear dynamic systems and its computational complexity, without loss of predictability.

- We have defined a formal quantitative language measure (Ray et.al. 2005), which is used to quantify statistical changes in the information space as we vary the operational setting of the network design space. We thus formulate theoretical foundations for solving the forward and backward problems of network adaptation by analytically associating a measure of the effect of changes in the network's topological structure to forecasts of system evolution (Fig. 10).

The actuation of network reconfiguration for large sensor networks is achieved through adaptive sampling at individual sensors, sensor mobility, turning existing sensors on or off, bandwidth reallocation, protocol modification, or through redeployment of resources. Urban topologies may further constrain such actuation, resulting in approximate solutions.

6.4 Conclusion
This research presents issues in the design of a distributed in-situ decision support system at the regional level that is capable of multi-level environmental monitoring and resource management using sensor networks. Diversity of sensor modalities is recommended for effective sensing, identification, and cross-cluster association of complex scenarios and environmental conditions. Analytical fusion models automate change detection and prediction as human oversight for low level sensing is not feasible in this amorphous networking environment.

6.5 Sensor network to determine drinking water quality and security
Finding patterns in large, real, spatial-temporal data continues to be of great interest. Ailamaki et al.(2003) describe a cross-disciplinary research effort to couple knowledge discovery in large environmental databases with biological and chemical sensor networks. They describe a distribution and operation protocol for the placement and utilization of in situ environment sensors by combining new algorithms for spatial-temporal data mining, new methods to model water quality and security dynamics, and a sophisticated analysis framework.

7. PLANS AND EXPECTED RESULTS FOR JALGAON DISTRICT
7.1 The District of Jalgaon, India

The objective of the program proposed in Jalgaon is to deploy a network of low cost, smart sensors to reduce or eliminate paper work, save field staff time and duplication efforts, improve the operational efficiency and accuracy of data, and provide timely data and reports to decision makers, like the District Collector, and to administrators, researchers, farmers, and the public. Jalgaon district is active in establishing digital governance process under the guidance of the District Collector, Mr. Vijay Singhal (MTech IIT) (http://jalgaon.nic.in/)

Just 20 years earlier, Jalgaon had a rich and healthy forest and ample natural resources. Back then the water table was just 80 feet deep – compared to the water table depth of 200 feet that is reality today. Over the last 20 years, the natural resources have been used in random and unmanaged ways, resulting in the current condition of resource scarcity, especially water. The “alarms have gone off” in Jalgaon, and this progressive district in the state of Maharashtra is eager to adopt modern sensor technology to better manage its natural resources. The district collector and the insightful district engineer V.D. Patil have initiated an innovative district level river linking program (http:www.jalgaon.nic.in/html/collmore.htm) to address water shortage problem.

The center of administration in Jalgaon is the District Collector’s office. This office was ISO 9000 certified in 2000, which demonstrates that the district is progressive and already has a well-defined process. There is a strong desire and eagerness; at all levels of the administration and with farmers, researchers, and the community, to aggressively improve the processes and implement more effective services for the people. Jalgaon has already adopted and implemented various modern tools and techniques to improve the workings of the administration. For example, it has implemented E-District, which is a district level website (http://jalgaon.nic.in/index.htm) where reports and data are made available to the public. Jalgaon now has the desire to use smart sensors to collect and digitize data for critical tasks, like infrastructure monitoring, healthcare development, disaster management, development projects, improved agricultural productivity, and water quality monitoring. The district level water resource management requires a multidisciplinary approach. Jalgaon district administration is working closely with JalaSRI, a research institute contributing significantly to the Surveillance and Research activities for Natural Resources Monitoring and Management for the Jalgaon District (http://www.jalasri.kces.in ). JalaSRI’s focus is on geo-informatics and hot spot detection under the leadership of Dr. G. P. Patil (http://www.stat.psu.edu/~gpp/). The needs and requirements of developing countries are very different than in developed countries. The availability of communications infrastructure and cost factors are unique to each country and to each district within the countries. The wireless sensor network and system design expertise for the Sensor Network Program is being provided by Erallo Technologies, Inc. USA (http://www.erallo.com ); which is active in research and development of wireless sensor networks and ad-hoc mesh networks.
Sensor Networks for Water Management

In the district of Jalgaon, water management is critical for drought conditions as well as for flood situations. 60% of the land in Jalgaon is classified as Drought Prone Areas (DPA). The remaining 40% has been classified as Assured Rain Fall Areas (ARFA); however, this classification was done many years ago and due to multiple factors, including global warming, they’ve seen an increase in the Drought Prone Areas (unofficial). This makes it even more critical for the government to provide effective water services and an efficient water management infrastructure to villages, farmers, and industry. The district of Jalgaon is a highly productive agricultural area; thus, the economy and politics of the area are a function of water availability.

In this area, water is managed using a series of interconnected dams: three major dams (the Girna, Hatnur and Waghur), 10 medium-sized dams, and 70 minor dams; along with 178 inspection-classified wells. Monitoring water availability and levels is a critical component for water management and irrigation projects. In the rainy season, measurements are taken every two to four hours (water depth, overflow, and amount of water). Based on these measurements, estimations are calculated for water over-flow predictions and the potential for down-stream flooding. In the dry season, the water depth is taken once a day. The measurements are used for irrigation and pumping rates and for drinking water availability. For example, when sufficient water is available, approximately 120 liters per person per day is made available; however, in drought conditions only 20 liters per person/day is provided. Thus, the accurate reading and reporting of the water depth is critical and fundamental for forecasting and decision making. With data from the sensor networks -- complete, concise, and real-time district level data on water entering the system (lakes, rivers, dam reservoir, and aquifers) will be used in a model to predict the availability of water within the district. This will be key information for the decision makers to formulate accurate water usage policies.

Today’s Problems with a Manual Process

Today, a manual process is used to record water depth in the dams. A technician takes a reading of the water depth (questionable accuracy) and calls it in by phone to the sub-division office. The sub-division office then calls in to the division level to report the measurement. The information is then written on a paper form at the division level. The set of reports for all the dams are sent to the Collector’ Office and then to the rest of the state wide agencies.

Inaccuracies in the data set can be introduced at multiple levels. The field measurements may be taken incorrectly or the low level field staff may not even take actual measurements and instead estimate the level. On numerous occasions, the data has been inconsistent and could not be correlated; however, since many days had passed by the time the inconsistency was found, it was too late to go back and re-measure or validate the levels. Another re-occurring problem is the time of day of the measurement -- which is frequently unreliable or not even provided. This makes the job of coordinating and synchronizing data with other dams or structures inaccurate or impossible. For example, a measurement reading scheduled to be taken at multiple structures (like dams or inspection wells) at 10:00 in the morning, may actually be read on one structure at 10:30 in the morning and in the evening at other structures.
A Solution for Water Management
Sensor devices could easily eliminate these problems. Most smart sensors (whether digital or analog) maintain an accurate clock; thus, the time-of-day that the data is collected is reported along with the measurement. Additionally, since sensors can be programmed to record measurements on a schedule or in response to a series of events, the data is always collected when required. Using cell phone technology, the sensor data can also be processed and automatically transmitted to a central office for further correlation and calculations -- in near real-time. Thus, the data at the district level as well as at the state level could be accurately and promptly processed and reported.

The district of Jalgaon is primarily an agro-economy – the entire agricultural planning for the district, including water shed management, management of drought prone areas, and resource scarcity management – are all heavily dependent on this critical set of data. A map of the Jalgaon district and its major and minor water harvesting projects are shown in Figure 10.

7.2 Sensors for Integrated Water Resource Management
An integrated network of various types of water sensors could provide a very comprehensive system for managing all the aspects of water resources: drinking and farming use; rain gauge measurements, dam depth and inspection well monitoring; water quality; planning for water harvesting, irrigation and pumping; and emergency responses for flood conditions. All of these aspects are interrelated – so then too should be the system to monitor them. Today, many of these tasks are not even routinely performed, further complicating the management of this growingly scarce resource. Measurements that are done come with very tedious manual and error-prone processes. Additionally, the time lag between data collection and reporting to authorities is too late to enable pro-active responses. Automated collection and reporting via sensors would improve the processes to provide this critical resource information on time and when needed.

Rain Fall Monitoring
Data from a sensor system to monitor rain fall would help officials anticipate the rain fall’s effect on the water table, water table recharge, and dams and reservoirs where water collects by providing the exact amount of water recharge due to rain. Sensors for monitoring dam reservoir depths and wells could be used to predict, manage, and prevent water overflow and flood potentials. More than just preventing floods, an integrated sensor network could also prevent the loss of a precious resource: rain water.

Currently, there are only 48 manually-read rain gauges spread throughout the district of Jalgaon. The measurements are taken at the village level and reported by phone to the tashil level (county level) at various times. Then, over 48 hours pass by before a district report is compiled and distributed to administrators, radio stations, and other authorities.

With the introduction of smart, low cost, wireless rain gauge sensors, hundreds of sensors could be deployed to provide the desired granularity at a micro level. The sensors would be programmed to accurately read the rainfall measurements, on a pre-defined schedule or in response to a pre-programmed series of events. At each sampling event, the rain gauge sensor...
would record the rain fall and time stamp of the reading. The data would also be automatically transmitted (using cell phone technology) to the district offices and/or state level offices – in real-time or near real-time -- on a regular schedule or in response to pre-defined events. A timely report could then be generated by software and stored in an open standards-based SQL database. By making the database structure public, any interested party could access the data within the Decision Support System (DSS).

**Rainfall Sensors**

A smart, electronic, digital rain sensor is designed to take accurate rainfall measurements from 0 to 300 mm per hour. The principle it uses is simple: water is collected by a funnel which is then routed to a measuring chamber. An electronic circuit is connected to the solid-state level sensor to output the measurement value into a programmable parameter value. When the capacity of the measuring chamber is met, the water is siphoned out (in about 1 second) onto the ground; and the process then repeats indefinitely. The electronics of the sensors are also programmed to filter out false readings, making the gauges very precise. The digital logger on the sensor can store the data for transmission. Many devices also include the time of day and outside temperature at the time of reading.

Rain monitoring sensor stations will be distributed throughout the district. The sensors will provide accurate hourly rain fall in millimeters, thus providing time series rain fall data and the rate of water storage recharge. With the time series measurement, the total amount of additional water in a given area can be accurately determined. The addition of water and rate data can also provide critical information on the rain’s effect on the aquifer as well as potential soil erosion damage. The sensors will be programmed to alter the sampling parameters once they detect rain and all sensor data will be sent to a central site; thus providing real-time data on rain and its distribution. The increase in aquifer levels could then be estimated and verified by sensors in the test wells. Time series data from each sensor would include the sensor ID, time stamp, and the amount of rain fall in millimeters/hour. Thus the data base would be populated with spatially distributed rainfall data. Temporal and spatial rainfall patterns could then be analyzed to describe the distribution of rail fall across the district.

**Well Monitoring**

The 178 inspection-class wells located throughout the Jalgaon district are used to monitor ground water levels. The depth of the water in the wells is manually measured once a month by the Groundwater Survey Development Agency (GSDA). The monthly data is compiled and a report is provided by the GSDA *once a year*. Thus, this critical data is not available to authorities on a timely basis. Further, the manual methods of measurement and recording the data are prone to error, skewing the report and any applications or calculations made from the data. The limited number and locations of the wells monitored also adds to the difficulty of providing a clear picture of the district. Additionally, in this agro-industrial area, the water table is greatly affected by excessive pumping, which in turn impacts farmers’ yields as well as precious drinking water. Improved monitoring of ground water discharge and recharge is critical.

An economical solution to inspection well monitoring is the use of low cost, wireless, smart sensors and software to automatically measure water depths, transmit the data, and generate
frequent reports. The use of sensors could also increase the number of inspection-class wells; which today is restricted by the number of technicians available to take and report the manual measurements.

The well monitoring sensor system will include district wide, spatially distributed, time series aquifer depth data -- enabling a district wide aquifer map. The long term analysis of the aquifer data will enable estimations and predictions on water availability in the future. The model could also be refined by using the farmer electricity usage as an indicator of aquifer discharge (by irrigation pumping).

Fig. 10: Map of the district of Jalgaon in the state of Maharashtra, India

Sensors for Well Water Depth
There are two main technologies used to measure water depth in wells and dams. The first type of sensor emits ultra sonic waves and then measures the time required for the waves to bounce back; similar to the sonar technique used by ships to measure the ocean depth. The sensors float on the surface of the water and send a ping at specified intervals. The time required for the wave to travel down and then bounce back is recorded and the water depth is accurately calculated. As these sensors float on the surface of the water, they are easily accessible for routine maintenance. The second technology uses piezo electric pressure sensors. These sensors are submersed to the bottom of the water body. From there they measure the water pressure on the sensor. The depth of the water is then extrapolated from the pressure exerted by the water. The sensor uses an
analog to digital converter to translate the analog signal into a digital format. The digital data is then processed and the pressure is converted to engineering units. Independent of the sensing technology used, all sensors provide spatial time series data on water levels in wells. If the well is equipped with an electric pump, the electricity usage of that pump will provide the well discharge rate. The amount of time required to recharge the well is an indicator of water availability in the aquifer.

**Reservoir and Dam Monitoring**

Monitoring is a very important task in appropriate dam management due to the economic, social and environmental significance of structures. It is fundamental in order to guarantee not only the safety of the structure and its users, but also to optimize the exploitation and maintenance of the dam.

Monitoring the water level in dams is critical for many water uses; including agricultural, community, and drinking water for both humans and livestock. Dam monitoring is also used to manage the water between dams to maintain their capacity and intactness. If up-to-date reports are made available on a timely basis, they could be used to predict water levels, better manage water availability, improve warning systems, and take corrective measures when necessary. Accuracy of the data is critical; any false data or false alarms causes excessive burdens on the first responders and wastes precious water.

Smart digital sensors are ideal for dam and reservoir monitoring. The automated sensors can sample the water level accurately and frequently so that up-to-date data is always available. The sensors can be configured to report on a regular schedule or report only special events, like the rate of change in water levels. Dam monitoring will be functionally categorized into two groups: water reservoir monitoring and dam safety monitoring. These two groups are interrelated and the database for both will be common. The sensor data will be mined in central data center. The hydrological model will extract data from the central database and generate customized reports for the dam operators and policy makers.

The Jalgaon district has complete cellular wireless coverage, thus, the backend connectivity can be economically achieved by using the existing cellular networks. A hard-wired phone line could be used as a backup system for data transmission. To eliminate false alarms, backup sensors will be deployed to continuously compare data with each other and generate an alarm only if all applicable sensors in the network agree with the event. A combination of fuzzy logic and voting algorithms will be implemented to reduce false alarm and identify/isolate any faulty sensors.
Sensors for Dam Monitoring

Commercial grade, water ultrasonic depth sensors (described above) are also ideal for dam and reservoir monitoring. Alternatively, the submersible piezo electric pressure sensors could also be used dam monitoring. These sensors measure the pressure of the surrounding water while subtracting the atmospheric pressure to determine an exact pressure exerted by a column of water. The depth of the water is then extrapolated from the pressure measurement. These sensors also have a dynamic temperature compensation system, enabling high accuracy measurements over a wide temperature range. The submersible pressure sensors have a solid state transducer which is encapsulated in a stainless steel housing and fully encapsulated with marine-grade epoxy to prevent moisture from leaking in.

One commercially available submersible water level sensor uses a unique, highly flexible silicon diaphragm to interface between the water and the sensing element. The silicon diaphragm protects the sensor's electronics from moisture, providing each sensor with reliable linearity and eliminating issues associated with metal foil diaphragms (which tend to crinkle and stretch out over time causing drift and linearity problems). This sensor also has automatic barometric compensation by using an attached vent cable which is protected by a stainless steel micro-screen cap to prevent fouling with silt, mud, or sludge.

Both ultrasonic and submersible sensors can be easily integrated with embedded processors, data loggers, and telemetry equipment. The processor would control the operations of the sensor as well as backend communication with the database. For applications in Jalgaon, the ideal backend connectivity may be GSM digital cellular.

Girna Dam Hydrologic Monitoring and Observing System (GDHMOS):

The primary objective of the Girna river sensor monitoring system is to design and implement a closed system to monitor and process water availability and usage of water in the Girna Dam water system. The term “closed system” refers to the scope of the system. The scope of the monitoring process will start with the rain fall into the river, its flow into the dam, and its passage all the way through the exit channels of the dam. The scope could also be expanded to take corrective actions based on sensor and user inputs. Although the heart of the system will be the Girna Dam and Girna Reservoir, sensing and monitoring will include recharging from the Girna River and remote monitoring of the rain fall in the recharging area, as well as, water discharge from the dam gates and monitoring of the rivers down stream.

The sensors will monitor the water volume (depth) of the reservoir, enable the calculation of hydraulic conductivity for dam safety, and keep track of the structural integrity of the dam. The sensors in the Girna River (the river feeding the reservoir) will include multi-layered, side-looking, Doppler current meters with variable range. These sensors can be easily installed on the bank of the river to measure the velocity of the water in horizontal layers so that accurate water flow can be measured. They can also be used to determine the volume of water entering the reservoir. Rain sensors deployed within a 5000 sq km recharge zone area (or catch basin) will monitor the amount of rain fall and provide a measurement of “anticipated additional water” expected to arrive at the reservoir and the anticipated time when the water will arrive to the reservoir. Sensors at the dam water discharge end (tow end) will measure the amount of water
leaving the Girna Dam system to the rivers/channels down stream and will enable an estimate of equilibrium or shift in equilibrium in the water level.

**Dam Safety Monitoring**
The second objective of GDHMOS is to monitor the safety of the dam. The sensor monitoring system will include safety sensors to monitor the structural integrity of the dam, as well as the hydraulic conductivity of the foundation and surrounding walls, to ensure dam safety. A wireless network of ultrasonic and piezo electric depth sensors installed at the foundation of the dam and in the inspection wells will continuously monitor the ground water flow under the dam. The dam design parameters and the constant monitoring of hydraulic conductivity will be used to determine the safety of the dam and avoid piping condition. The structural safety of the dam will be monitored by a network of crack measuring sensors, strain gauges, and fiber optic foundation deformation sensors. Any minute changes in the dam structure could be quickly identified and reported. The sensor system will help ensure that the dam is operating under safe conditions and provide information to anticipate any adverse events within given conditions; like weather, water volume, discharge rate, cracks in the dam, and seepage variation.

With the set of data discussed above, engineers and policy makers could better assess, estimate, and plan water flow and usage – and the effects it may have on people down stream from the dam. For example, early warnings could be issued when the dam gates must be opened to eliminate excessive water. People living in a ‘danger zone’ could be put on alert, and with pre-planning, the water could be released gradually. Thus, any disruption downstream could be effectively eliminated or managed.

**Intelligent Sensors and Software**
The data from the sensors will be sampled at rates specified by the user or sampled on event-based conditions. In an event-driven situation, the sensors would detect an abnormal situation and then automatically alter the sampling rate, thereby providing data a higher resolution. The heart of the monitoring system will be specialized intelligent sensors that can react to the environmental conditions and central data mining software capable of performing analysis and compiling reports. The report distribution facility will include computerized custom reports.

**Agricultural, Forest and Weather Monitoring**
The economy of the Jalgaon district is based on agriculture and the agro industry in the region – which makes the accurate monitoring of rain fall and water levels for agricultural irrigation and food processing even more critical. Monitoring needs to be on both a micro and macro level and is the foundation for related applications, like ground moisture, disease outbreaks or forecasts, and pest control. For instance, cropping patterns are based on the prevailing rainfall as measured through the current rain gauge methods and the GSDA inspection wells (which are often inaccurate and out-of-date). The forest department in this region also depends on this critical (yet often flawed today) data to manage its resources.
Sensors for Agro/Forest Monitoring
Moisture sensors can play a vital role in helping the farmer determine planting patterns for gram and rabbi crops. In addition to aiding the farmers on the most advantageous sowing times, other benefits include improved planning for seed availability and fertilizers at appropriate times and places, depending on the moisture condition at specific locations. In addition to the moisture sensors for the soil, improved methods for monitoring rainfall, dams, wells, and irrigation reservoirs with sensor networks would be a great service to the farming community and economy of the Jalgaon district. This information could easily be distributed to farmers and forest managers.

Sensors for Micro-weather Station and Early Warning System
A typical weather station includes anemometer sensors to measure wind speed and wind direction; electronic rain gauges; sensors for barometric pressure changes; and humidity sensors for relative humidity, heat index, and dew point. Micro-weather station sensors can provide timely and critical weather measurement data. For instance, a plunging barometer suggests an incoming storm and an anemometer sensor can capture threateningly-high wind gusts or dangerously low wind chill conditions.

A chain of micro weather stations could be established in the district and integrated with the water management network. This would allow the establishment of a robust early warning system. These sensors can be integrated with an embedded controller to locally process the sensor data and issue warnings/alarms if the data presented by the sensor exceeds normal threshold. The data collection and reporting scheme could also be remotely monitored and changed in real-time. The location of each sensor node would be identified by its sensor address. As this would be a static network, the location of each sensor would be pre-determined, enabling DSS software to conduct analyses based on the geographic location of each sensor. Since each sensor in the sensor network would have a unique address, each sensor could be individually addressed and configured. This configuration could be done locally or remotely from a central unit. The remote connectivity and the local microprocessor based controller would provide two way command and control capabilities to each sensor within a sensor node and within the sensor network.

A sensor package could be customized based on the regional needs. Each weather sensor node would contain one or more analog or digital sensors. Each individual sensor would be controlled by a microprocessor responsible for recording and transmitting the data. The software in the controller could also be configured to apply intelligent knowledge and generate critical alarms if the weather sensor readings deviate from a normal, pre-defined range.

7.3 In Conclusion on Sensors and Expected Results
The efficiency of a digital governance system for policy decisions and availability of data to the public could be significantly improved with the use of smart wireless sensors nodes and sensor networks. In a dynamic environment, data becomes ineffective if it is not processed as soon as it is received. Sensors and sensor networks will provide timely data and make the decision support systems used by policy makers more robust; include emergency and disaster responses, forecasts and planning, and resource management of water, irrigation, and forests. The services of the administration could be better focused and tuned towards the needs of the community and people and the administration could more quickly respond to existing and emerging hot-spots because
of the availability of fast and accurate data. Thus, the use of sensors and sensor networks in government operations would all greatly benefit the regional public, agricultural industry, and commerce.

8. FUTURISTIC VISION

8.1 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. Information 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.

8.2 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. Graduate students will test and validate various tools as they become available through the project. Also, the graduate students will 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: monographs, case books, thematic journal issues, research workshops and tutorials, and distributed information management.

Acknowledgements:
This material is based upon work partially supported by (1) The National Science Foundation under Grant No.0307010 (ii) The US Environmental Protection Agency under Grant No. CR - 83059301 and No. R-828684-01; and (iii) The US Army Research Office Award W911NF-07-0376.
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 sponsoring agencies.
D. REFERENCES


Fishburn, P. C. Interval Orders and Interval Graphs: A Study of Partially Ordered Sets.


JaJa, J. and Shi, Q. Efficient Techniques for Exploring Geospatial Data, 2001


http://www.stat.psu.edu/~gpp/pdfs/SyndromicSurveillance Comment.pdf

http://www.stat.psu.edu/~gpp/PDFs/TR2002-0203.pdf

http://www.stat.psu.edu/~gpp/pdfs/TR2002-0501.pdf

http://www.stat.psu.edu/~gpp/pdfs/Poster 1.pdf

http://www.stat.psu.edu/~gpp/pdfs/TR2002-0601.pdf


