

Upper Level Set Scan Statistic System for Detecting Arbitrarily Shaped Hotspots for Digital Governance*

G. P. Patil and S. L. Rathbun
 Penn State University
 Department of Statistics
 University Park, PA 16802 USA
 Tel.: 001 814 865 9442
 Email: gpp@stat.psu.edu

R. Acharya and P. Patankar
 Penn State University
 Dept of Computer Sci & Engineering
 University Park, PA 16802 USA
 Tel: 001 814 865 9505
 Email: acharya@cse.psu.edu

Reza Modarres
 George Washington University
 Department of Statistics
 Washington, DC 20052 USA
 Tel: 001 202 994 6359
 Email: reza@gwu.edu

ABSTRACT

A declared need is around for geoinformatic surveillance statistical science and software infrastructure for spatial and spatiotemporal hotspot detection. Hotspot means something unusual, anomaly, aberration, outbreak, elevated cluster, critical resource area, etc. The declared need may be for monitoring, etiology, management, or early warning. The responsible factors may be natural, accidental, or intentional. This paper suggests methods and tools for hotspot detection across geographic regions and across networks. The investigation proposes development of statistical methods and tools that have immediate potential for use in critical societal areas, such as public health and disease surveillance, ecosystem health, water resources and water services, transportation networks, persistent poverty typologies and trajectories, environmental justice, biosurveillance and biosecurity, censor networks, robotics, video mining, social networks, and others. We introduce, for multidisciplinary use, an innovation of the health-area-popular circle-based spatial and spatiotemporal scan statistic. Our innovation employs the notion of an upper level set, and is accordingly called the upper level set scan statistic, pointing to a sophisticated analytical and computational system as the next generation of the present day popular SaTScan. Success of surveillance rests on potential elevated cluster detection capability. But the clusters can be of any shape, and cannot be captured only by circles. This is likely to give more of false alarms and more of false sense of security. What we need is capability to detect arbitrarily shaped clusters. The proposed upper level set scan statistic innovation is expected to fill this need. This five year NSF DGP project has been instrumental to conceptualize surveillance geoinformatics partnership among several interested cross-disciplinary scientists in academia, agencies, and private sector. The planned poster is expected to reveal several live case studies and outcomes of real geospatial and spatiotemporal data sets of current interest.

Categories and Subject Descriptors

H. Information systems; H4. Information systems applications; H4.2 types of systems.

General Terms

Decision support for hotspot detection, prioritization, and early warning.

Keywords

Confidence set of hotspots, early warning, geosurveillance statistics, hotspot detection, hotspot rating, typology of space-time hotspots, surveillance geoinformatics partnership.

1. INTRODUCTION

In response to an ever increasing volume of georeferenced data, government agencies require a new generation of decision support systems for early detection, surveillance, and prioritization of hotspots. Hotspots are unusual phenomena, anomalies, aberrations, outbreaks, or critical areas. Government agencies require hotspot delineation and prioritization for etiology, management, and early warning.

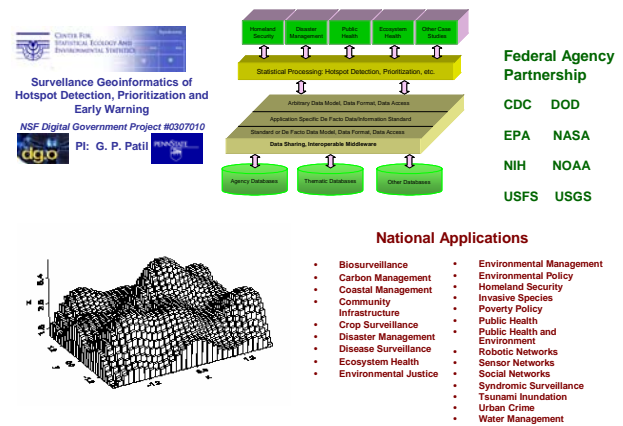


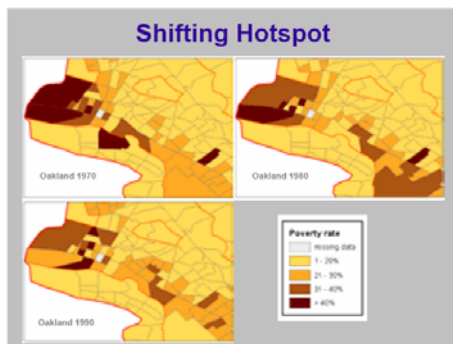
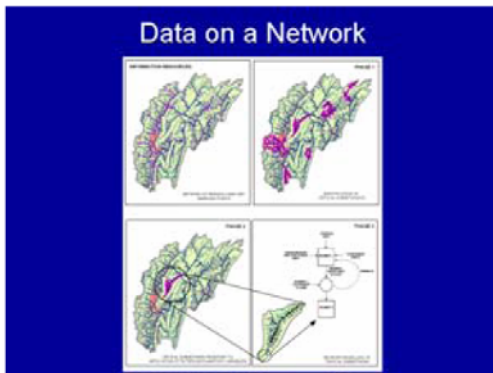
Figure 1. NSF Digital Government surveillance geoinformatics project, federal agency partnership and national applications for digital governance.

* This paper is dedicated to Charles Taillie, our longtime friend and versatile colleague.

With support from the NSF/DG program, we have developed a decision support framework for geographic and network surveillance. Our framework will be illustrated in a number of case studies of potential interest to several federal agencies.

2. UPPER LEVEL SET SCAN STATISTIC SYSTEM

Our framework features a novel *upper level set scan statistic* (ULS) system to delineate arbitrarily shaped hotspots in both spatial and temporal dimensions [1]. This approach can be applied to irregular networks, such as those formed by streams (see below), political units, social networks, and the internet. When applied to data collected over both space and time, the ULS scan statistic system can be used to detect shifting hotspots (see below), coalescence of neighboring hotspots, or their growth.



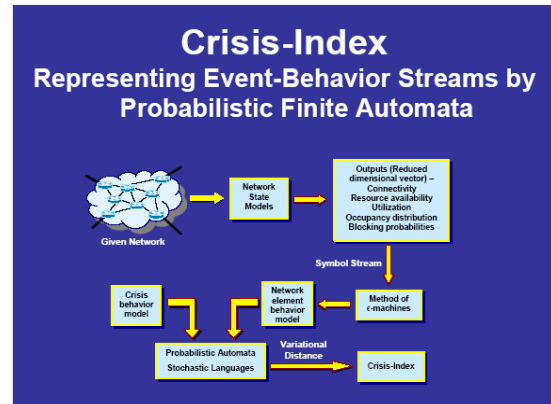
3. PARTIALLY ORDERED SET PRIORITIZATION SYSTEM

We also propose a novel prioritization scheme based on multiple indicators that does not require the reduction of the data to a single index. This *poset prioritization and ranking system* features Haase diagrams describing the partial ordering of data, linear extension decision trees illustrating admissible rankings among hotspots, and cumulative rank functions for hotspot prioritization [2].

4. SENSOR NETWORKS

Geotelemetry employs self-healing wireless networks of smart sensors, each receiving and communicating high dimensional data streams. We propose a *probabilistic finite state automaton* (PFSA), describing a network element as obtained from its data stream output. The variational distance between the stochastic

languages generated by normal and crisis automata is used to form a crisis index. The ULS scan statistic is then applied to crises indices over a collection of network elements for hotspot detection. These hotspots and their prioritization can be used for objects identification and their trajectories. Additional applications of PFSA include the tasking of self-organizing surveillance mobile sensor networks, videomining networks, syndromic surveillance in public health, and habitat monitoring in ecosystem health.



Distance Between Two PFA

Let A and B be two PFAs on the same alphabet Σ

Let $w(i)$ be a probability distribution across string lengths i

Let π_A and π_B be the w -weighted probability measures of A and B

Define the distance between A and B as the **variational distance** between the probability measures π_A and π_B :

$$d(A, B) = \|\pi_A - \pi_B\|$$

For additional information regarding our project, see <http://www.stat.psu.edu/hotspots/> and <http://www.stat.psu.edu/~gpp/>

5. ACKNOWLEDGMENTS

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, (ii) the United States Environmental Protection Agency under Grant No. CR-83059301 and (iii) the Pennsylvania Department of Health using Tobacco Settlement Funds under Grant No. ME 01324. 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.

6. REFERENCES

- [1] Patil, G.P., and Taillie, C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11 (2004), 183-197.
- [2] Patil, G.P., and Taillie, C. Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics*, 11 (2004), 199-228.