Surveillance GeoInformatics

Hotspot Detection, Prioritization, and Early Warning

G. P. Patil

December 2004 – January 2005
Geoinformatic spatio-temporal data from a variety of data products and data sources with agencies, academia, and industry.

Masks, filters

Indicators, weights

Spatially distributed response variables

Hotspot analysis

Prioritization

Geoinformatic Surveillance System

Decision support systems

Masks, filters
The Spatial Scan Statistic

- Move a circular window across the map.
- Use a variable circle radius, from zero up to a maximum where 50 percent of the population is included.
A small sample of the circles used
Detecting Emerging Clusters

- Instead of a circular window in two dimensions, we use a cylindrical window in three dimensions.
- The base of the cylinder represents space, while the height represents time.
- The cylinder is flexible in its circular base and starting date, but we only consider those cylinders that reach all the way to the end of the study period. Hence, we are only considering ‘alive’ clusters.
West Nile Virus Surveillance in New York City

- 2000 Data: Simulation/Testing of Prospective Surveillance System
- 2001 Data: Real Time Implementation of Daily Prospective Surveillance
West Nile Virus Surveillance in New York City

Major epicenter on Staten Island

- Dead bird surveillance system: June 14
- Positive bird report: July 16 (coll. July 5)
- Positive mosquito trap: July 24 (coll. July 7)
- Human case report: July 28 (onset July 20)
Voice response telephone reporting system

Website dead bird report form

Dead bird reports

SQL Server
dead bird report database

ESRI geocoding

SAS
data management

SaTSan cluster analysis

ESRI mapping

Policy-makers
Hospital Emergency Admissions in New York City

- Hospital emergency admissions data from a majority of New York City hospitals.
- At midnight, hospitals report last 24 hour of data to New York City Department of Health
- A spatial scan statistic analysis is performed every morning
- If an alarm, a local investigation is conducted
Issues

Limitations and Needs

- Circles capture only compactly shaped clusters
- Want to identify clusters of arbitrary shape
- Circles handle only synoptic (tessellated) data
- Want to also handle data on a network
- Circles provide point estimate of hotspot
- Want to assess estimation uncertainty (hotspot confidence set)

Poor Hotspot Delineation by Space-Time Cylinders

Poor Hotspot Delineation by Circular Zones

[Diagram showing space-time cylinders and circular zones with annotations]

Cylindrical approximation sees single hotspot as multiple hotspots
Geospatial Surveillance

Synoptic Data
Hotspots within Hotspots

Data on a River Network

Estimation Uncertainty in Hotspot Identification
Spatial Temporal Surveillance

### Shifting Hotspot

- Oakland 1870
- Oakland 1880
- Poverty rate:
  - missing data
  - 1 - 20%
  - 21 - 40%
  - 41 - 60%
  - > 40%

### Typology of Space-Time Hotspots

- Stationary Hotspot
- Expanding Hotspot
- Shifting Hotspot
- Merging Hotspot

### Trajectory of a Merging Hotspot

1970
1980
1990
2000
Syndromic Crisis-Index Surveillance

**Crisis-Index**
Representing Event-Behavior Streams by Probabilistic Finite Automata

Probabilistic Finite Automata (PFA)

A PFA is a DFA \((Q, q_0, \Sigma, \delta)\) with a probability attached to each transition such that the sum of the probabilities across all transitions from a given node is unity.

Formally, \(p: Q \times \Sigma \rightarrow [0, 1]\) such that
- \(p(q, a) = 0\) if and only if \(\delta(q, a) = \text{Blocked}\)
- \(\sum_{a \in \Sigma} p(q, a) = 1\) for all \(q \in Q\)

Multiplying branch probabilities lets us assign a probability value \(\mu(q_0, s)\) to each string \(s\) in \(\Sigma^*\). E.G., \(\mu(q_0, abcd) = (0.8)(1.0)(0.4) = 0.32\)

**Distance Between Two PFA**

Let \(A\) and \(B\) be two PFA's on the same alphabet \(\Sigma\)

Let \(\nu(i)\) be a probability distribution across string lengths \(i\)

Let \(\pi_A\) and \(\pi_B\) be the \(\nu\)-weighted probability measures of \(A\) and \(B\)

Define the distance between \(A\) and \(B\) as the variational distance between the probability measures \(\pi_A\) and \(\pi_B\):

\[d(A, B) = \| \pi_A - \pi_B \|\]
Hotspot Prioritization

Prioritization of Disease Clusters with Multiple Indicators

Data Matrix

<table>
<thead>
<tr>
<th>cluster</th>
<th>SIR</th>
<th>LL</th>
<th>Young Cases</th>
<th>Multiple Cancers</th>
<th>Atypical Demographics</th>
<th>Late Stage of Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF2</td>
<td>2.09</td>
<td>10.36</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LM14</td>
<td>1.5</td>
<td>36</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>LM4</td>
<td>2.04</td>
<td>19.21</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>LF7</td>
<td>1.51</td>
<td>15.43</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B2</td>
<td>1.21</td>
<td>31.3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>B4</td>
<td>1.25</td>
<td>28.4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LM1</td>
<td>2.32</td>
<td>21.91</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LM3</td>
<td>2.13</td>
<td>21.26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LM7</td>
<td>2.12</td>
<td>13.33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

* LF = lung, female; LM = lung, male; B = breast

Ranking Partially Ordered Sets

Cumulative Rank Frequency Operator

The curves are stacked one above the other giving a linear ordering of the elements: a > b > c > d > e > f
Surveillance Geoinformatics and Digital Governance

An NSF Digital Government Project

Welcome to our website

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, 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.
National Applications

- Biosurveillance
- Carbon Management
- Coastal Management
- Community Infrastructure
- Crop Surveillance
- Disaster Management
- Disease Surveillance
- Ecosystem Health
- Environmental Justice
- Sensor Networks
- Robotic Networks
- Environmental Management
- Environmental Policy
- Homeland Security
- Invasive Species
- Poverty Policy
- Public Health
- Public Health and Environment
- Syndromic Surveillance
- Social Networks
- Stream Networks
Securing the nation's computer networks from cyber attack is an important aspect of Homeland Security. Project develops diagnostic tools for detecting security attacks, infrastructure failures, and other operational aberrations of computer networks.

Many critical applications of surveillance sensor networks involve finding hotspots. The upper level set scan statistic is used to guide the search by estimating the location of hotspots based on the data previously taken by the surveillance network.

New York City has installed 892 drinking water sampling stations. Currently, about 47,000 water samples are analyzed annually. The ULS scan statistic will provide a real-time surveillance system for evaluating water quality across the distribution system.

Emerging hotspots for disease or biological agents are identified by modeling events at local hospitals. A time-dependent crisis index is determined for each hospital in a network. The crisis index is used for hotspot detection by scan statistic methods.

West Nile virus is a serious mosquito-borne disease. The mosquito vector bites both humans and birds. Scan statistical detection of dead bird clusters provides an early crisis warning and allows targeted public education and increased mosquito control.

Disruption of American agriculture and our food system could be catastrophic to the nation's stability. This project has the specific aim of developing novel remote sensing methods and statistical tools for the early detection of crop bioterrorism.

The scan statistic hotspot delineation and poset prioritization tools will be used in combination with our oil spill detection algorithm to provide for early warning and spatial-temporal monitoring of marine oil spills and their consequences.

This study employs 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.
Hotspot Detection Innovation
Upper Level Set Scan Statistic

Attractive Features

- Identifies arbitrarily shaped clusters
- Data-adaptive zonation of candidate hotspots
- Applicable to data on a network
- Provides both a point estimate as well as a confidence set for the hotspot
- Uses hotspot-membership rating to map hotspot boundary uncertainty
- Computationally efficient
- Applicable to both discrete and continuous syndromic responses
- Identifies arbitrarily shaped clusters in the spatial-temporal domain
- Provides a typology of space-time hotspots with discriminatory surveillance potential
Candidate Zones for Hotspots

- **Goal**: Identify geographic zone(s) in which a response is significantly elevated relative to the rest of a region.
- A list of candidate zones $Z$ is specified *a priori*
  - This list becomes part of the parameter space and the zone must be estimated from within this list.
  - Each candidate zone should generally be spatially connected, e.g., a union of contiguous spatial units or cells.
  - Longer lists of candidate zones are usually preferable.
  - Expanding circles or ellipses about specified centers are a common method of generating the list.
Scan Statistic Zonation for Circles and Space-Time Cylinders

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

Outbreak expanding in time
- Small cylinders miss much of the outbreak
- Large cylinders include many unwanted cells
Question: Are there data-driven (rather than *a priori*) ways of selecting the list of candidate zones?

Motivation for the question: A human being can look at a map and quickly determine a reasonable set of candidate zones and eliminate many other zones as obviously uninteresting. Can the computer do the same thing?

A data-driven proposal: Candidate zones are the connected components of the upper level sets of the response surface. The candidate zones have a tree structure (echelon tree is a subtree), which may assist in automated detection of multiple, but geographically separate, elevated zones.

Null distribution: If the list is data-driven (i.e., random), its variability must be accounted for in the null distribution. A new list must be developed for each simulated data set.
ULS Scan Statistic

- Data-adaptive approach to reduced parameter space $\Omega_0$
- Zones in $\Omega_0$ are **connected components** of **upper level sets** of the empirical intensity function $G_a = \frac{Y_a}{A_a}$
- Upper level set (ULS) at level $g$ consists of all cells $a$ where $G_a \geq g$
- Upper level sets may be disconnected. Connected components are the candidate zones in $\Omega_0$
- These connected components form a rooted tree under set inclusion.
  - Root node = entire region $R$
  - Leaf nodes = local maxima of empirical intensity surface
  - Junction nodes occur when connectivity of ULS changes with falling intensity level
Upper Level Set (ULS) of Intensity Surface

Intensities $g$ on Region $R$

Hotspot zones at level $g$
(Connected Components of upper level set)
Changing Connectivity of ULS as Level Drops
N.B. Intensity surface is cellular (piece-wise constant), with only finitely many levels. A, B, C are junction nodes where multiple zones coalesce into a single zone.
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.
Network Analysis of Biological Integrity in Freshwater Streams
New York City Water Distribution Network

Water Sampling Station
- Nearly 1000 sampling stations throughout New York City
- Over 47,000 samples were collected and tested in 2001

Water Testing and Treatment Facilities
- Continuous monitoring of water quality flowing into tunnels and aqueducts
- Water treatment and disinfection
NYC Drinking Water Quality
Within-City Sampling Stations

• 892 sampling stations
• Each station about 4.5 feet high and draws water from a nearby water main
• Sampling frequency increased after 9-11
Currently, about 47,000 water samples analyzed annually
• Parameters analyzed:
  - Bacteria
  - Chlorine levels
  - pH
  - Inorganic and organic pollutants
  - Color, turbidity, odor
  - Many others
Network-Based Surveillance

- Subway system surveillance
- Drinking water distribution system surveillance
- Stream and river system surveillance
- Postal System Surveillance
- Road transport surveillance
- Syndromic Surveillance
Syndromic Surveillance

- Symptoms of disease such as diarrhea, respiratory problems, headache, etc.
- Earlier reporting than diagnosed disease.
- Less specific, more noise.
Syndromic Surveillance

(Left) The overall procedure, leading from admissions records to the crisis index for a hospital. The hotspot detection algorithm is then applied to the crisis index values defined over the hospital network.

(Right) The $\varepsilon$-machine procedure for converting an event stream into a parse tree and finally into a probabilistic finite state automaton (PFSA).
Experimental Validation

Pressure sensitive floor

Formal Language Events:
  a – green to red or red to green
  b – green to tan or tan to green
  c – green to blue or blue to green
  d – red to tan or tan to red
  e – blue to red or red to blue
  f – blue to tan or tan to blue

Wall following

Random walk

Target Behavior

Analyze String Rejections
Emergent Surveillance Plexus (ESP)
Surveillance Sensor Network Testbed
Autonomous Ocean Sampling Network
Types of Hotspots

- Hotspots due to multiple, localized, stationary sources
- Hotspots corresponding to areas of interest in a stationary mapped field
- Time-dependent, localized hotspots
- Hotspots due to moving point sources
Ocean Sampling Mobile Network
OSAMON
Ocean SAMplng MOBILE Network OSAMON Feedback Loop

- Network sensors gather preliminary data
- ULS scan statistic uses available data to estimate hotspot
- Network controller directs sensor vehicles to new locations
- Updated data is fed into ULS scan statistic system
SAmpling MOBILE Networks (SAMON)

Additional Application Contexts

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

Mote, Smart Dust: Small, flexible, low-cost sensor node

Miniaturized Spec Node Prototype

RF Component of Alcohol Sensor

Giner’s Transdermal Alcohol Sensor
Scalable Wireless Geo-Telemetry with Miniature Smart Sensors

Geo-telemetry enabled sensor nodes deployed by a UAV into a wireless ad hoc mesh network: Transmitting data and coordinates to TASS and GIS support systems.
Architectural Block Diagram of Geo-Telemetry Enabled Sensor Node with Mesh Network Capability
Standards Based Geo-Processing Model
UAV Capable of Aerial Survey
Data Fusion Hierarchy for Smart Sensor Network with Scalable Wireless Geo-Telemetry Capability
Wireless Sensor Networks for Habitat Monitoring
Target Tracking in Distributed Sensor Networks
Video Surveillance and Data Streams
Video Surveillance and Data Streams
Turning Video into Information
Measuring Behavior by Segments

- Customer Intelligence
- Enterprise Intelligence
- Entrance Intelligence
- Media Intelligence
- Video Mining Service
Deterministic Finite Automata (DFA)

Directed Graph (loops & multiple edges permitted) such that:

- Nodes are called States
- Edges are called Transitions
- Distinguished initial (or starting) state
- Transitions are labeled by symbols from a given finite alphabet, $\Sigma = \{a, b, c, \ldots\}$
- The same symbol can label several transitions
- A given symbol can label at most one transition from a given state (deterministic)
Deterministic Finite Automata (DFA)

Formal Definition

Quadruple \((Q, q_0, \Sigma, \delta)\) such that:

- \(Q\) is a finite set of states
- \(\Sigma\) is a finite set of symbols, called the alphabet
- \(q_0 \in Q\) is the initial state
- \(\delta: Q \times \Sigma \rightarrow Q \cup \{\text{Blocked}\}\) is the transition function:
  - \(\delta(q, a) = \text{Blocked}\) if there is no transition from \(q\) labeled by \(a\)
  - \(\delta(q, a) = q'\) if \(a\) is a transition from \(q\) to \(q'\)
Any path through the graph starting from the initial state determines a string from the alphabet.

**Example:** The blue dashed path determines the string \( a \ b \ c \ a \)

Conversely, any string from the alphabet is either **blocked** or determines a path through the graph.

**Example:** The following strings are blocked:
\[ c, \ aa, \ ac, \ abb, \ \text{etc.} \]

**Example:** The following strings are not blocked:
\[ a, \ b, \ ab, \ bb, \ \text{etc.} \]

The collection of all unblocked strings is called the **language** accepted or determined by the DFA (all states are “final” in our approach)
Strings and Languages

\( \Sigma = \) (finite) alphabet
\( \Sigma^* = \) set of all (finite) strings from \( \Sigma \)

A language is any subset of \( \Sigma^* \).

Not all languages can be determined by a DFA. Different DFAs can accept the same language

Let \( \Sigma^i = \Sigma \times \Sigma \times \ldots \times \Sigma \) \((i\)-fold cartesian product\).
\( \Sigma^i \) consists of all strings of length \( i \).

Then, \( \Sigma^* \) decomposes as

\[ \Sigma^* = \bigcup_{i=1}^{\infty} \Sigma^i = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \ldots \]
A PFA is a DFA \((Q, q_0, \Sigma, \delta)\) with a probability attached to each transition such that the sum of the probabilities across all transitions from a given node is unity.

Formally, \(p: Q \times \Sigma \rightarrow [0, 1]\) such that

- \(p(q, a) = 0\) if and only if \(\delta(q, a) = \text{Blocked}\)
- \(\sum_{a \in \Sigma} p(q, a) = 1\) for all \(q \in Q\)

Multiplying branch probabilities lets us assign a probability value \(\mu(q_0, s)\) to each string \(s\) in \(\Sigma^*\). E.G., \(\mu(q_0, abca) = (0.8)(0.6)(0.4) = 0.192\)
Properties of $\mu(q_0, s)$

- For fixed $q_0$, $\mu(q_0, s)$ is a **measure** on $\Sigma^*$
- **Support** of $\mu$ is the language accepted by the DFA
- For fixed $q_0$, $\mu(q_0, s)$ is a **probability measure** on $\Sigma^i$
  \[ (\Sigma^i = \text{strings of length } i) \]
  This probability measure is written as $\mu^{(i)}$.
- Given a probability distribution $w(i)$ across string lengths $i$,
  \[
  \pi(q_0, s) = \sum_{i=0}^{\infty} w(i) \mu^{(i)}(q_0, s)
  \]
  defines a **probability measure** across $\Sigma^*$, called the $w$-weighted probability measure of the PFA.
If all $w(i)$ are positive, then the support of $\pi$ is also the language accepted by the underlying DFA.
Distance Between Two PFA

Let $A$ and $B$ be two PFAs on the same alphabet $\Sigma$

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

Let $\pi_A$ and $\pi_B$ be the $w$-weighted probability measures of $A$ and $B$

Define the distance between $A$ and $B$ as the \textit{variational distance} between the probability measures $\pi_A$ and $\pi_B$:

$$d(A, B) = \| \pi_A - \pi_B \|$$
Crop Attack Decision Support System

Site Identification Module

- Crops
- Key Crop Areas
- NOAA Weather
- Threat Locations

Signature Development Module

- Plants
  - Infected
  - Non-infected
  - Sentinel
- Ground Cameras
- Air/Space Platforms

Hyperspectral Imagery

Data Processing

Signature Library

Anomaly Report

Ground Truthing

Crop Attack Decision Support System

Site Identification Module

- Crops
- Key Crop Areas
- NOAA Weather
- Threat Locations

Signature Development Module

- Plants
  - Infected
  - Non-infected
  - Sentinel
- Ground Cameras
- Air/Space Platforms

Hyperspectral Imagery

Data Processing

Signature Library

Anomaly Report

Ground Truthing
Crop Biosurveillance/Biosecurity

Spectral Signature Deformation

Healthy Plant

Diseased Plant

Biological Agent

$T$ Photosynthesis Mapping Operator

$T : V \rightarrow W$

Spectral Mapping

$V = \text{Healthy Signature}$

$W = \text{Diseased Signature}$
Crop Biosurveillance/Biosecurity Data Processing Module

- Hyperspectral Imagery
- Image Segmentation (hyperclustering)
- Proxy Signal (per segment)
- Disease Signature
- Similarity Index (per segment)
- Tessellation (segmentation) of raster grid
- Signature Similarity Map
- Hotspot/Anomaly Detection
We also present a prioritization innovation. It lies in the ability for prioritization and ranking of hotspots based on multiple indicator and stakeholder criteria without having to integrate indicators into an index, using Hasse diagrams and partial order sets. This leads us to early warning systems, and also to the selection of investigational areas.
**HUMAN ENVIRONMENT INTERFACE**
**LAND, AIR, WATER INDICATORS**

for land - % of undomesticated land, i.e., total land area-domesticated (permanent crops and pastures, built up areas, roads, etc.)
for air - % of renewable energy resources, i.e., hydro, solar, wind, geothermal
for water - % of population with access to safe drinking water

<table>
<thead>
<tr>
<th>RANK</th>
<th>COUNTRY</th>
<th>LAND</th>
<th>AIR</th>
<th>WATER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sweden</td>
<td>69.01</td>
<td>35.24</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Finland</td>
<td>76.46</td>
<td>19.05</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>Norway</td>
<td>27.38</td>
<td>63.98</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>Iceland</td>
<td>1.79</td>
<td>80.25</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>Austria</td>
<td>40.57</td>
<td>29.85</td>
<td>100</td>
</tr>
<tr>
<td>22</td>
<td>Switzerland</td>
<td>30.17</td>
<td>28.10</td>
<td>100</td>
</tr>
<tr>
<td>39</td>
<td>Spain</td>
<td>32.63</td>
<td>7.74</td>
<td>100</td>
</tr>
<tr>
<td>45</td>
<td>France</td>
<td>28.34</td>
<td>6.50</td>
<td>100</td>
</tr>
<tr>
<td>47</td>
<td>Germany</td>
<td>32.56</td>
<td>2.10</td>
<td>100</td>
</tr>
<tr>
<td>51</td>
<td>Portugal</td>
<td>34.62</td>
<td>14.29</td>
<td>82</td>
</tr>
<tr>
<td>52</td>
<td>Italy</td>
<td>23.35</td>
<td>6.89</td>
<td>100</td>
</tr>
<tr>
<td>59</td>
<td>Greece</td>
<td>21.59</td>
<td>3.20</td>
<td>98</td>
</tr>
<tr>
<td>61</td>
<td>Belgium</td>
<td>21.84</td>
<td>0.00</td>
<td>100</td>
</tr>
<tr>
<td>64</td>
<td>Netherlands</td>
<td>19.43</td>
<td>1.07</td>
<td>100</td>
</tr>
<tr>
<td>77</td>
<td>Denmark</td>
<td>9.83</td>
<td>5.04</td>
<td>100</td>
</tr>
<tr>
<td>78</td>
<td>United Kingdom</td>
<td>12.64</td>
<td>1.13</td>
<td>100</td>
</tr>
<tr>
<td>81</td>
<td>Ireland</td>
<td>9.25</td>
<td>1.99</td>
<td>100</td>
</tr>
</tbody>
</table>
Hasse Diagram
(all countries)
Hasse Diagram
(Western Europe)
Ranking Partially Ordered Sets – 5

Poset (Hasse Diagram)

Linear extension decision tree

Jump Size: 1 3 3 2 3 5 4 3 3 2 4 3 4 4 2 2
Cumulative Rank Frequency Operator – 5
An Example of the Procedure

In the example from the preceding slide, there are a total of 16 linear extensions, giving the following cumulative frequency table.

<table>
<thead>
<tr>
<th>Element</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>9</td>
</tr>
<tr>
<td>b</td>
<td>7</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
</tr>
</tbody>
</table>

Each entry gives the number of linear extensions in which the element (row label) receives a rank equal to or better that the column heading.
The curves are stacked one above the other and the result is a linear ordering of the elements: \( a > b > c > d > e > f \)
Cumulative Rank Frequency Operator – 7
An example where $F$ must be iterated

Original Poset (Hasse Diagram)

$F^2$
Incorporating Judgment
Poset Cumulative Rank Frequency Approach

• Certain of the indicators may be deemed more important than the others

• Such differential importance can be accommodated by the poset cumulative rank frequency approach

• Instead of the uniform distribution on the set of linear extensions, we may use an appropriately weighted probability distribution $\pi$, e.g.,

$$\pi(\omega) = w_0 + w_1 n_1(\omega) + w_2 n_2(\omega) + \cdots + w_p n_p(\omega)$$
Figure 2 - All nodes within 1 step [direct link] of original suspects
Figure 3 - All Nodes within 2 steps / degrees of original suspects
Space-Time Poverty Hotspot Typology

- Federal Anti-Poverty Programs have had little success in eradicating pockets of persistent poverty

- Can spatial-temporal patterns of poverty hotspots provide clues to the causes of poverty and lead to improved location-specific anti-poverty policy?
Covariate Adjustment

Known Covariate Effects (age, population size, etc.)

\[ Y_a = \text{count in cell } a \]
\[ Y_a \sim \text{Poisson}(\lambda_a A_a) \text{ where } \lambda_a = \text{unknown relative risk for cell } a \]
\[ A_a = \text{known numerical covariate adjustment} \]

Hotspot Hypothesis Testing Model

\[ H_0 : \lambda_a \text{ are equal for all cells } a \text{ (constant relative risk)} \]
\[ H_1 : \lambda_a \text{ take two distinct values, an elevated value in an unknown zone } Z \text{ and a smaller value outside } Z \]

List of candidate zones \( Z \) (ULS approach)

All connected components of upper level sets

of the adjusted cellular surface \( Y_a / A_a \)
Covariate Adjustment

Given Covariates, Unknown Effects

\[ Y_a \sim \text{Poisson}(\lambda_a A_a) \quad \text{where} \quad \lambda_a = \text{unknown relative risk} \quad \text{for cell } a \]
\[ A_a = \text{unknown covariate adjustment} \]

GLM Model

\[ X_a = \text{vector of known covariate values for cell } a \]
\[ \beta = \text{vector of unknown covariate effects} \]

Model: \[ \log(A_a) = X_a^T \beta \quad \text{or} \quad \log(\lambda_a A_a) = \eta_a = \log(\lambda_a) + X_a^T \beta \]

Hotspot Hypothesis Testing Model

\[ H_0 : \lambda_a \text{ are equal for all cells } a \text{ (constant relative risk)} \]
\[ H_1 : \lambda_a \text{ take two distinct values, an elevated value in an unknown zone } Z \]
\[ \text{and a smaller value outside } Z \]

List of candidate zones \( Z \) (ULS approach)

All connected components of upper level sets of the adjusted cellular surface \( Y_a / A_a \)

Here the model must be fitted under the null hypothesis before determining the adjustments \( A_a \) and the candidate zones \( Z \)
Incorporating Spatial Autocorrelation

Ignoring autocorrelation typically results in:
- **under**-assessment of variability
- **over**-assessment of significance ($H_0$ rejected too frequently)

How can we account for possible autocorrelation?

**GLMM (SAR) Model**

$Y_a = \text{count in cell } a$  \hspace{1cm} $Y_a$ distributed as Poisson  \hspace{1cm} $\eta_a = \log(\text{E}[Y_a])$

The $Y_a$ are **conditionally** independent given the $\eta_a$

The $\eta_a$ are jointly Gaussian with a Simultaneous AutoRegressive (SAR) specification

$$\eta_a - \mu_a = \rho \sum_{b} W_{ab} (\eta_a - \mu_a) + \varepsilon_a$$

Here,  $\mu_a = E[\eta_a]$

$\varepsilon_a$ are iid $N(0, \sigma^2)$

$W_{ab}$ is a spatial weight expressing the "degree of association"

between cells $a$ and $b$ (Take $W_{aa} = 0$ and $W_{ab} = 1$)

Thus, the residual $\eta_a - \mu_a$ for cell $a$ is a **deflated** (by $\rho$) weighted average of the residuals for neighboring cells plus a disturbance term $\varepsilon_a$
Incorporating Spatial Autocorrelation

SAR Model: \[ \eta_a - \mu_a = \rho \sum_b W_{ab} (\eta_a - \mu_a) + \varepsilon_a \]

Matrix Form: \[ \eta - \mu = \rho W (\eta - \mu) + \varepsilon \]
\[ \eta - \mu = (I - \rho W)^{-1} \varepsilon \]
\[ \eta \sim \text{MVN} \left( \mu, \sigma^2 \left[ (I - \rho W)^T (I - \rho W) \right]^{-1} \right) \]

Unknown Parameters: \( \mu_a, \rho, \sigma^2 \)

Special Cases: \( \sigma^2 = 0 \Rightarrow \text{classical (iid) spatial scan} \)
\( (\rho \text{ is not identifiable here}) \)
\( \rho = 0, \sigma^2 \neq 0 \Rightarrow \text{overdispersed classical scan} \)
Incorporating Spatial Autocorrelation

GLMM (SAR) Model

\[ Y_a \sim \text{Poisson}(\exp(\eta_a)) = \text{Poisson}(\exp(\mu_a) \exp(\eta_a - \mu_a)) = \text{Poisson}(\lambda_a A_a) \]

where \( \eta \sim \text{MVN}\left( \mu, \sigma^2 \left[ (I - \rho W)^T (I - \rho W) \right]^{-1} \right) \)

Hotspot Hypothesis Testing Model

\( H_0 : \mu_a \) are equal (to \( \mu \)) for all cells \( a \) (constant relative risk)

\( H_1 : \mu_a \) take two distinct values, an elevated value in an unknown zone \( Z \) and a smaller value outside \( Z \)

List of candidate zones \( Z \) (ULS approach)

All connected components of upper level sets

of the \text{adjusted} cellular surface \( Y_a / A_a \) where \( A_a = \exp(\eta_a - \mu) \)

Here the model must be fitted under the null hypothesis (\( \mu_a = \mu \)) before determining the adjustments \( A_a \) and the candidate zones \( Z \)
Spatial Autocorrelation Plus Covariates

In the SAR model, \( Y_a \sim \text{Poisson}(\exp(\eta_a)) \)

where \( \eta \sim \text{MVN} \left( E[\eta], \sigma^2 \left[ (I - \rho W)^T (I - \rho W) \right]^{-1} \right) \),

express the mean of \( \eta \) as

\[ E[\eta] = \mu + X\beta \]

and formulate the \textit{Hotspot Hypothesis Testing Model} in terms of the constant term \( \mu \) in this expression.
CAR Model

The entire formulation is similar for Conditional AutoRegressive (CAR) specs except that the form of the variance-covariance matrix of $\eta$ is changes.

In the CAR model, $Y_a \sim \text{Poisson}(\exp(\eta_a))$ where

$$\eta \sim \text{MVN} \left( E[\eta], \sigma^2 (I - \rho W^*)^{-1} A^* \right)$$

and $A^*$ is diagonal.

However, parameters in CAR and SAR have very different interpretations. In CAR, the conditional variances are

$$\text{Var}(\eta_a | \eta_b, b \neq a) = \sigma^2 A_{aa}^*$$

which (strangely) do not depend on the autocorrelation parameter $\rho$.

In SAR, the conditional variances are

$$\text{Var}(\eta_a | \eta_b, b \neq a) = \sigma^2 / (1 + \rho^2 \sum_b W_{ba}^2)$$

This expression is intuitively appealing since the conditional variances are decreasing functions of $\rho^2$ and are smallest for cells $a$ with many strongly associated neighbors (relatively large $W_{ba}^2$ for many $b$).
Spatially distributed response variables

Hotspot analysis

Prioritization

Decision support systems

Geoinformatic Surveillance System

Geoinformatic spatio-temporal data from a variety of data products and data sources with agencies, academia, and industry

Masks, filters

Indicators, weights

Masks, filters