

Chapter 9

DEFENDING AGAINST CATASTROPHIC TERRORISM

Chapter Overview

Terrorist attacks can cause devastating damage to a society through the use of chemical, biological, or radiological weapons. Biological attacks may cause contamination, infectious disease outbreaks, and significant loss of life. Information systems that can efficiently and effectively collect, access, analyze, and report infectious disease data can help prevent, detect, respond to, and manage these attacks. In this chapter we discuss our ongoing BioPortal research and system development efforts motivated to address some of these challenges.

9.1 Case Study 14: BioPortal for Disease and Bioagent Surveillance

Our initial BioPortal research focuses on two prominent infectious diseases: *West Nile Virus* (WNV) and *Botulism* (BOT). These two diseases were chosen because of their significant public health and national security implications and the availability of related datasets with our research partners. We developed a research prototype, called the *WNV-BOT Portal* system, which provides integrated, web-enabled access to a variety of distributed data sources including the New York State Department of Health (NYSDOH), the California Department of Health Services (CADHS), and some other federal sources (e.g., United States Geological Survey, USGS). It also provides advanced information visualization capabilities as well as predictive modeling support.

Architecturally, the WNV-BOT Portal consists of three major components: a *web portal*, a *data store*, and a *communications backbone*. The web portal component implements the user interface and provides the following main functionalities: (1) searching and querying available WNV/BOT datasets, (2) visualizing WNV/BOT datasets using spatial-temporal visualization, (3) accessing analysis and prediction functions, and (4) accessing the alerting mechanism.

To enable data interoperability, we use Health Level Seven (HL7) standards (<http://www.hl7.org/>) as the main storage format. In our data warehousing approach, contributing data providers transmit data to WNV-BOT Portal as HL7-compliant XML messages (through a secure network connection if necessary). To alleviate potential computational performance problems associated with this HL7 XML-based approach, we have identified a core set of data fields based on which searches could be performed efficiently.

An important function of the data store layer is data ingest and access control. The data ingest control module is responsible for checking the integrity and authenticity of data feeds from the underlying information sources. The access control module is responsible for granting and restricting user access to sensitive data.

The communication backbone component enables data exchanges between the WNV-BOT Portal and the underlying WNV/BOT sources based upon the CDC's National Electronic Disease Surveillance System (NEDSS) and HL7 standards. It uses a collection of source-specific "connectors" to communicate with underlying sources. We use the connector linking NYSDOH's Health Information Network (HIN) system and WNV-BOT Portal to illustrate a typical design of such connectors. The data from HIN to the portal system is transmitted in a "push" manner. HIN sends secure Public

Health Information Network Messaging System (PHIN MS) messages to the portal at pre-specified time intervals. The connector at the portal side runs a data receiver daemon listening for incoming messages. After a message is received, the connector checks for data integrity syntactically and invokes the data normalization subroutine. Then the connector stores the verified message in the portal's internal data store through its data ingest control module. Other data sources (e.g., those from USGS) may have "pull"-type connectors which periodically download information from the source web sites and examine and store data in the portal's internal data store.

The WNV-BOT Portal makes available the Spatial Temporal Visualizer (STV) (Buetow et al., 2003) to facilitate exploration of infectious disease case data and to summarize query results. The STV has three integrated and synchronized views: periodic, timeline, and GIS. Figure 9-1 illustrates how these three views can be used to explore an infectious disease dataset. The top-left panel shows the GIS view. The user can select multiple datasets to be shown on the map in a layered manner using the checkboxes. The top-right panel corresponds to the timeline view displaying the occurrences of various cases using a Gantt chart-like display. The user can also access case details easily using the tree display located left of the timeline display. Below the timeline view is the periodic view through which the user can identify periodic temporal patterns (e.g., which months have an unusually high number of cases). The bottom portion of the interface allows the user to specify subsets of data to be displayed and analyzed.

Our project has supported exploration of, and experimentation with, technological infrastructures needed for the full-fledged implementation of a national infectious disease information infrastructure and has helped foster information sharing and collaboration among related government agencies at state and federal levels. In addition, we have obtained important insights and hands-on experience with various important policy-related challenges faced by developing a national infrastructure. For example, a significant part of our project activity has been centered on developing privacy-sensitive data sharing agreements between project partners from different states.

Our ongoing technical research is focusing on two aspects of infectious disease informatics: hotspot analysis and efficient alerting and dissemination. For WNV, localized clusters of dead birds typically identify high-risk disease areas. Automatic detection of dead bird clusters using hotspot analysis can help predict disease outbreaks and allocate prevention/control resources effectively. Initial experimental results indicate that these techniques are promising for disease informatics analysis. We are planning to augment existing predictive models by considering additional environmental factors (e.g., weather information, bird migration patterns) and tailoring data mining techniques for infectious disease datasets that have prominent temporal features.

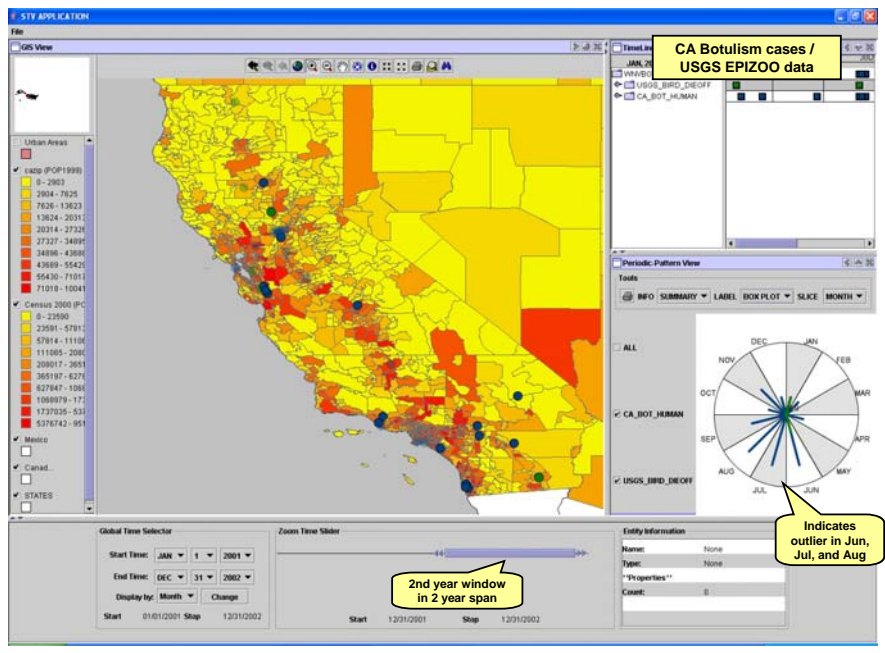


Figure 9-1. Using STV to visualize botulism data.

9.2 Case Study 15: Hotspot Analysis and Surveillance

In infectious disease informatics and bioterrorism studies, measurements of interest are often made at various locations and with timestamps. In the public health domain, the disease cases reported to the Centers for Disease Control and Prevention (CDC) through its National Notifiable Diseases Surveillance System are collected with timestamps at various places across the entire nation. Similar disease case reporting practices exist at state and local jurisdictions, typically with cases identified with specific geo-locations.

Recent years have seen increasing interest in answering the following central questions of great practical importance arising in spatio-temporal data analysis and related predictive modeling: (a) How can we identify areas having exceptionally high or low measures? (b) How can we determine whether the unusual measures can be attributed to known random variations or are statistically significant? In the latter case, how do we assess the explanatory factors? (c) How can we identify any statistically significant changes (e.g., in rates of health syndromes or crime occurrences) in a timely manner in a geographic area?

Two types of approaches have been developed in the literature to address some of these questions. The first type of approach is considered the

retrospective model. It is aimed at testing statistically whether a disease is randomly distributed over space and time for a predefined geographical region during a predetermined time period. The second type of approach is *prospective* in nature, with repeated time periodic analyses targeted at identification of statistically significant changes in real time. Both approaches have been applied in security informatics practice. For instance, the New York City Department of Health and Mental Hygiene collected geo-coded information concerning dead birds infected by West Nile Virus (WNV). Applying retrospective methods to avian fatality data, they were able to detect possible WNV outbreaks before human case data became available (Zeng et al., 2004).

Our study focuses on retrospective models. Although a wide range of methods have been proposed for retrospective spatio-temporal data analysis, the space scan statistic, in particular, has become one of the most popular methods for detection of disease clusters and is being widely used by many public health departments and researchers. Algorithmically, the space scan statistic method imposes a circular window on the map under study and moves the center of the circle over the area so that at different positions the window includes different sets of neighboring cases. Over the course of data analysis, the method creates a large number of distinct circular windows (other shapes such as rectangle and ellipse have also been used), each with a different set of neighboring areas within it and each a possible candidate for containing a cluster of events. Conditioning on the observed total number of cases, the spatial scan statistic is defined as the maximum likelihood ratio over all possible circles. The likelihood ratio for a circle indicates how likely the observed data are given a differential rate of events within and outside the zone.

Despite its success, there are a number of limitations associated with the scan statistic approach. First, its efficiency depends on the use of simple, fixed symmetrical shapes of regions. As a result, when the real underlying clusters do not conform to such shapes, the identified regions are often not well localized. Second, it is difficult to customize and fine-tune the clustering results using the scan statistic approach. Users often have different needs as to the level of granularity and number of the resulting clusters and they may have different degrees of tolerance regarding outliers.

These limitations have motivated our research aimed at exploring and developing alternative and complementary modeling approaches for spatio-temporal data analysis in the context of security informatics. This case study reports our effort in studying two such approaches: *Risk-adjusted Nearest Neighbor Hierarchical Clustering* (RNNH), which was initially developed for crime analysis; and *Risk-adjusted Support Vector Clustering* (RSVC), a new hotspot analysis approach we have recently developed.

Developed for crime hotspot analysis, RNNH is based on the well-known nearest neighbor hierarchical clustering (NNH) method, combining the hierarchical clustering capabilities with kernel density interpolation techniques. The standard NNH approach identifies clusters of data points that are close together (based on a threshold distance). Many such clusters, however, are due to some background or baseline factors (e.g., the population which is not evenly distributed over the entire area of interest). RNNH is primarily motivated to identify clusters of data points *relative* to the baseline factor. Algorithmically, it dynamically adjusts the threshold distance inversely proportional to some density measure of the baseline factor (e.g., the threshold should be shorter in regions where the population is high). Such density measures are computed using kernel density based on the distances between the location under study and some or all other data points.

RSVC is the result of our recent attempt to combine the risk adjustment idea of RNNH with a modern, robust clustering mechanism such as Support Vector Machines (SVM) to improve the quality of hotspot analysis. SVMs are the most well-known of a class of algorithms that use the idea of kernel substitution. SVM-based data description and novelty detection (DDND) is particularly relevant to our research. SVM-based DDND methods are aimed at identifying the *support* of a data distribution. They can single out data clusters in complex shapes and have been well-tested in complex, noisy domains (e.g., handwritten symbol recognition). The standard version of SVM-based DDND does not take into consideration baseline data points and therefore cannot be directly used in spatio-temporal data analysis. As such, we have developed a *risk-adjusted* variation, called RSVC, based on ideas similar to those in RNNH. In this new approach we first compute the kernel density estimations using the baseline data points and then adjust width parameter in the Gaussian kernel function based on such density estimations.

We have conducted a series of computational studies to evaluate the effectiveness of the three hotspot analysis techniques (Satscan, RNNH, RSVC) discussed above. In our experiment, we used artificially generated datasets with known underlying probability distributions to precisely and quantitatively evaluate the efficacy of these techniques. We use the well-known measures from information retrieval to evaluate the performance of hotspot techniques: Precision, Recall, and F-Measure. In the spatial data analysis context, we define these measures as follows.

Let A denote the size of the hotspot(s) identified by a given algorithm, B the size of the true hotspot(s), and C the size of the overlapped area between the algorithm-identified hotspot(s) and true hotspot(s). Precision is defined as C/A . Recall is defined as C/B . F-measure is defined as the harmonic mean of precision and recall ($2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$). High

recall indicates low false negatives and high precision indicates low false positives. Notice that achieving a high level of one measure often impacts negatively achieving a high level of the other measure. F-measure represents a balance and trade-off between precision and recall.

We report three artificially-generated scenarios with which we have experimented. In the first scenario, as shown in Figure 9-2, we first randomly generated 100 baseline points in one circle (whose center is around (6.7, 8.2) on the graph).

We then generated 200 case points of interest in total. The first 100 points were generated inside of the baseline circle. (As such, these data points do not represent a new unusual hotspot.) The next 100 points were generated within another circle (whose center is around (14.3, 10.2)). It is clear that the only true hotspot is this second circle.

To make the problem more interesting, we introduced some noise -- 30 outlier baseline points and 40 outlier case points over the entire map. For statistical testing purposes, we repeated the above data generation process 30 times to produce 30 instances of the template scenario by moving the centers of the two circles randomly across the map and also by varying the distances between them so long as they did not overlap.

To collect the performance data, we ran all hotspot analysis methods under study on all the problem instances. Table 9-1 summarizes these methods' average performance across all instances. Overall RSVC achieves the best F-score at 84.5%.

Table 9-1. Average performance of RSVC, SCAN, and RNNH on 30 instances of Scenario 1.

Technique	Precision	Recall	F-measure
RSVC	79.5%	92.4%	84.5%
SCAN	54.3%	92.9%	65.4%
RNNH	95.3%	49.0%	64.0%

Table 9-2. Average performance of RSVC, SCAN, and RNNH on 30 instances of Scenario 2.

Technique	Precision	Recall	F-measure
RSVC	78.5%	72.6%	75.0%
SCAN	60.0%	77.4%	67.4%
RNNH	87.9%	42.3%	56.2%

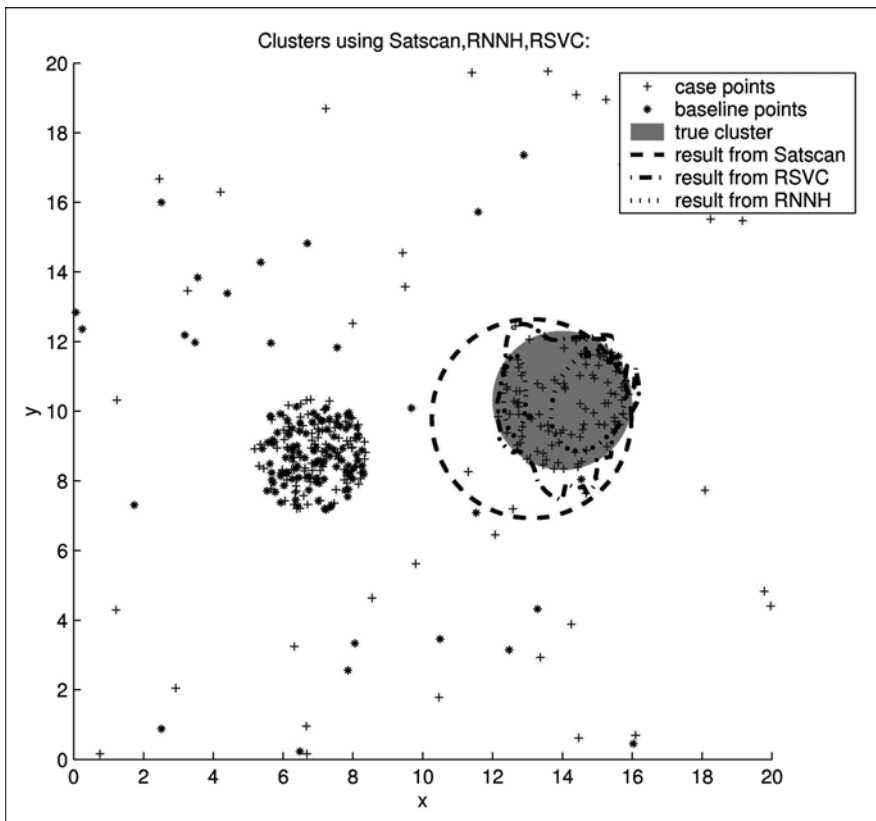


Figure 9-2. Scenario 1 based on simulated data.

Two additional scenarios have been investigated in our study. In Scenario 2, the true hotspots are the two pieces left of a rectangle when its middle section is removed by a circle. In Scenario 3, the true spot is a square with its circular-shaped center removed. The purpose behind the use of these scenarios is to test how robust these hotspot methods are when dealing with hotspots of irregular shape. Tables 9-2 and 9-3 summarize the average performance of these methods when applied to identify hotspots in 30 randomly-generated instances of Scenario 2 and Scenario 3, respectively.

Table 9-3. Average performance of RSVC, SCAN, and RNNH on 30 instances of Scenario 3.

Technique	Precision	Recall	F-Measure
RSVC	84.3%	71.6%	77.2%
SCAN	78.7%	65.0%	69.9%
RNNH	77.0%	24.6%	36.6%

Some observations are in order. In general, RSVC and the spatial scan method have similar levels of recall across different scenarios. However, RSVC has higher precision than the spatial scan method (confirmed by statistical tests). RNNH has the highest precision level but typically with the lowest recall. When considering the combined F-measure, RSVC consistently delivers the best results. For complex, irregular shapes, RSVC typically performs the best among the three techniques. We believe RSVC could be a strong candidate for hotspot identification in national security applications (e.g., disease surveillance, crime hotspot analysis).

9.3 Future Directions

Over the past ten years we have witnessed the many devastating global impacts of infectious diseases such as SARS, foot-and-mouth disease, avian influenza, West Nile Virus, etc. Many of these diseases have strong implications for national security (i.e., as bioagents). Disease and bioagent surveillance is one of the most critical research topics of relevance to homeland security.

Traditionally, disease informatics research is conducted by epidemiologists in universities and local, state, and federal public health agencies. However, there are many opportunities for advanced information sharing, retrieval, and visualization research which could adapt some of the new digital library, search engine, and information visualization techniques. In addition, the discipline can benefit greatly from new temporal and spatial data mining techniques that could complement the prevailing statistical analysis techniques in epidemiology. Standards and ontologies are also critically needed in disease informatics research.

9.4 Questions for Discussion

1. What are the prevailing techniques for disease and syndromic surveillance in epidemiology? What are the assumptions?
2. What is the current status of the BioWatch program at DHS and the BioSense program at CDC? What are their strengths and deficiencies? What are some research opportunities?
3. What are some ways to work with local and state public health agencies on diseases informatics research? What are the available datasets?
4. How should the data confidentiality issue of animal and human disease information be addressed?

5. How can social network analysis and hotspot analysis techniques be applied to other types of diseases? How can other predictive and environmental modeling techniques be incorporated?

