SAND REPORT

SAND2002-3747
Unlimited Release
Printed November 2002

Syndrome Surveillance using Parametric Space-Time Clustering

Mark W. Koch, Sean A. McKenna and Roger L. Bilisoly

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2002-3747
Unlimited Release
Printed November 2002

Syndrome Surveillance using Parametric Space-Time Clustering

Mark W. Koch
Signal and Image Processing Systems Department

Sean A. McKenna and Roger L. Bilisoly
Geohydrology Department

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0844

Abstract

As demonstrated by the anthrax attack through the United States mail, people infected by the biological agent itself will give the first indication of a bioterror attack. Thus, a distributed information system that can rapidly and efficiently gather and analyze public health data would aid epidemiologists in detecting and characterizing emerging diseases, including bioterror attacks.

We propose using clusters of adverse health events in space and time to detect possible bioterror attacks. Space-time clusters can indicate exposure to infectious diseases or localized exposure to toxins. Most space-time clustering approaches require individual patient data. To protect the patient's privacy, we have extended these approaches to aggregated data and have embedded this extension in a sequential probability ratio test (SPRT) framework. The real-time and sequential nature of health data makes the SPRT an ideal candidate. The result of space-time clustering gives the statistical significance of a cluster at every location in the surveillance area and can be thought of as a "health-index" of the people living in this area.

As a surrogate to bioterrorism data, we have experimented with two flu data sets. For both databases, we show that space-time clustering can detect a flu epidemic up to 21 to 28 days earlier than a conventional periodic regression technique. We have also tested using simulated anthrax attack data on top of a respiratory illness diagnostic category. Results show we do very well at detecting an attack as early as the second or third day after infected people start becoming severely symptomatic.

Contents

Figures

Tables

1.0 Introduction

Early detection, identification, and warning are essential to minimize casualties from a biological attack. Ongoing efforts within Departments of Energy and Defense focus on developing and deploying environmental sensor systems. However, sensors cannot cover all potential targets, and attacks may be vector-based rather than environmentally based. Therefore, sick people will likely provide the first indication of an attack. An enhanced medical surveillance system that detects an attack as soon as people become symptomatic could dramatically increase the lives saved.

The ultimate surveillance objective would be a distributed information system for gathering and then analyzing public health data in near-real time, so that epidemiologists can detect and characterize emerging diseases, including bioterror attacks, more rapidly than ever before. The primary collection points for data would consist of the health-care providers who will, in the near future, document patient records electronically. With additional new tools, other researchers can draw on the same data for epidemiological and medical outcome research giving a wide range of additional benefits for the health of Americans.

1.1. Problem and Approach

The problem is to use on-line electronic medical information to detect and characterize a biological weapons attack after only a relatively small number of days (1-2) have passed from when the victims first begin to present symptoms. To sift through the massive amount of medical records and focus attention on potentially suspicious places and times, we propose to use space-time clustering. Space-time clustering searches for statistically significant clusters of adverse health events in space and time, can indicate exposure to infectious diseases or localized exposure to toxins, and help pinpoint the location of the contaminant's source. See (Williams 1984) for an excellent review of space-time clustering tests of individual case histories.

Clustering has a number of other advantages. Unlike supervised approaches, clustering algorithms belong to the class of unsupervised algorithms and don't need data on bioterror events to train the algorithm. Until the recent anthrax attack through the United States (US) mail, there have been only two published bioterror attacks on US soil. One happened in Oregon in 1984 when the Bhagwan Shree Rajneesh religious cult tried to influence a local political election by introducing salmonella into salad bars, sickening 751 people (Torok 1997). Another happened in 1996 at a Texas medical lab when a disgruntled worker tainted muffins and doughnuts with shigella bacteria causing dysentery (Kolavic 1997). Even if these attacks were sufficient to train an algorithm, there was no surveillance system to capture the data. One could try to simulate data representing a bioterror attack to use as training data, but this is not without its problems. One runs the risk of having an algorithm key on irregularities in the simulation that may not occur in real life. Extremely detailed simulations not only take a long time to run, but require many runs for a classification algorithm to successfully generalize to the full range of possible variables.

Clustering also has a multiple use characteristic. Space-time anomalies not only arise from bioterror attacks, but new and emerging infectious diseases or epidemics can also lead to space-time clusters. This dual use goal not only allows one to justify the cost, but also to prove that the algorithms really work. For example, we will show we can detect influenza epidemics 21-28 days earlier than conventional periodic regression techniques.

2.0 Detecting a Statistically Significant Cluster

Besides finding space-time clusters we need to determine their statistical significance or understand the errors of making a decision of whether a cluster exists or not. Figure 2-1 shows a simple example of having incidence data from one region in time and trying to determine if the maximum incidence (indicated by the dot) represents an anomaly or not. As you go down, each graph shows the maximum number of cases increasing. The 200 data samples are randomly generated using a Poisson distribution with a mean of 2. One approach for detecting the anomaly sets an arbitrary threshold and hopes for the best. The best approach statistically determines the errors that could occur from a making decision. Obviously we don't want to set the threshold so low that we have a false alarm every day. Here, our faith in the system would quickly degrade. We also don't want to set the threshold so high that the system would miss true anomalies.

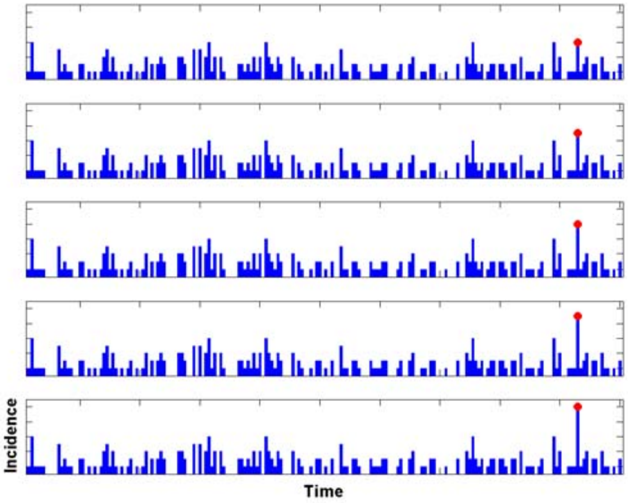

Figure 2-1. Monte Carlo Simulation of Incidence in Time for one Region.

The two errors that can occur are called type I and type II errors. If H_0 represents the hypothesis that the observed data comes from the background and H_1 the hypothesis that the data comes from an anomaly, then the type I error defines the false alarm (FA) error and the type II error defines the missed detection error. Here, false alarms result from calling a normal background event an anomaly and missed detections result from calling an anomaly a background event. Often, we can only reduce one error at the expense of increasing the other error. For example, to reduce the MD error the FA error will increase. To reduce both errors we need to improve the quality of the data or incorporate data from other sources. As in space-time clustering, other sources of data can come from neighboring regions or the incidence at previous times. While we have yet to incorporate disparate sources in our work such as 911 calls or over-the-counter drug sales, our framework allows incorporation of these sources.

If we understand the stochastic process that generates the background time series then we can set a threshold to control the FA errors or the P-value. For the simple example in Figure 2-1, the P-value represents the probability of getting the maximum value or higher in the background noise. The probability of getting a maximum value *z* in *n* independent samples from a Poisson distribution with mean λ is:

$$\mathrm{p}_{\mathrm{val}} = 1 - \left(\sum_{i=0}^{z} \frac{\lambda^i e^{-\lambda}}{i!} \right)^n \quad \textbf{(2-1)}$$

Note this requires knowledge that the underlying background process comes from a Poisson distribution and estimation of the parameter λ. Only when the P-value dips below a small threshold do we want to sound the alarm.

Figure 2-2 shows the same graphs as before, but now we've add P-values. You probably wouldn't want to declare an anomaly until the P-value was 9/1000 or 2/1000. Of course it depends of severity of the disease and how many false alarms you can tolerate.

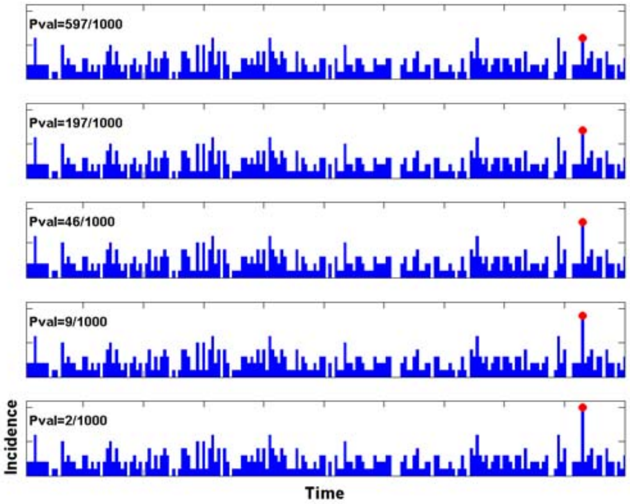

Figure 2-2. Same Data as Figure 2-1 with P-values Added.

It's more difficult to control the MD errors since this requires knowledge of the stochastic process that generates the anomaly. As we will discuss later, one can use historical data, for example epidemic data, to understand the stochastic anomaly process or one can use a form of power analysis (Murphy 1998, Cohen 1988). Power analysis assumes knowledge of the effect or anomaly size, which allows estimation of the power or the probability of detecting the anomaly. Understanding the effect size can come from experience or simulation.

For space-time clustering with real data, the underlying process may vary from data set to data set, and we also will have hundreds or thousands of regions each with their own time series. To reduce the errors we want to combine data across various time and space scales, but correlations in time and space reduce the amount of "information" available. In the rest of this report we will concentrate on these problems and their solutions.

3.0 Previous Work on Space-time Clustering

3.1. Data Types

In reviewing previous work on space-time clustering we have found it convenient to divide space-time data into two types:

1. Point data
2. Aggregate data.

Point data record the exact time and place of where the symptoms occurred and most often arise in the investigation of a public health concern. For example, a member of the public will report a large number of cases of a chronic disease. They might want to know if these cases are larger than expected, or if a cause is still operating and endangering the public. A public health official can interview the people with the chronic disease and determine the addresses of where each spends the most time and the times of when the symptoms appeared. This approach uses a number of individual case histories, each of which is considered to be a single data point.

Aggregated data combine the space-time information in a window of space and time, and are most often found in publicly available research data sets. For example, instead of reporting a person's home address, we can report just their zipcode, county, state, or country. Time of the disease's appearance can be aggregated into days, weeks, months, quarters, or years. Aggregation can reduce the size of the data. Aggregated data can also insure the privacy of a patient. We can get aggregated data from point data, but not vice versa. While aggregation is convenient for the reasons cited, valuable information is often lost. Due to privacy concerns we expect syndrome surveillance systems to aggregate data.

3.2. Clustering in Space and Time

Popular space-time statistical tests like Knox (Knox 1964a and 1964b), Mantel (Mantel 1967), and *k*-Nearest Neighbor (Jacquez 1996) require point-type data. See (Williams 1984) for an excellent review of space-time clustering tests. It's possible to threshold the adverse health event incidence to get a substitute to point cases. We have found this approach to require data from a large number of regions in order to estimate the background distribution of the test (McKenna 2002).

3.3. Clustering in Time Only

Most time clustering tests use data with the results aggregated within a time window. The simplest test assumes the aggregated samples come from a Poisson distribution (Stroup 1989 and Knox 1982). Whereas this Poisson test uses disjoint windows, the scan test uses overlapping windows and allows for the fixed-length aggregation-window to scan along a time series searching for the interval with the maximum number of cases (Weinstock 1981 and Wallenstein 1980). These tests use distributions determined by Naus (Naus 1965 and 1966). Tango has developed a clustering index that weights the distance between time intervals (Tango 1984). Chen uses the time between significant events (Chen 1978). In the real data sets we have investigated, we have not found evidence of a Poisson distribution. These wrong assumptions can lead to inaccurate estimates of statistical significance of the clusters.

3.4. Clustering in Space Only

Most space-clustering tests measure spatial auto-correlation between regions and stem from tests developed by Geary (Geary 1954) and Moran (Moran 1948 and 1950). Variations on these involve a binary transform of the data (i.e. significant versus not significant) (Moran 1948, Grimson 1981, Hungerford 1991) and using *join* statistics (Cliff 1981). Other variations include using weighted distances (Cliff 1981, Lai 1997, Walter 1994), and transforming the data to its rank (Cliff 1981, Lai 1997, Walter 1994). Combinations of the above possibilities lead to more variations.

In space clustering, many ways exist for determining the distance between the regions. One could use adjacency of regions, distance between region centroids, minimum distance between major cities for each region, amount of boundary shared by each region, or distance based on airline routes and major thoroughfares. Other choices involve how the contagion might spread and use distances based on population size, population density, or distance from largest outbreak.

One problem with clustering based on spatial-correlation is when the test measures correlation between non-significant events. For example, an event in a "sea" of non-significant events would result in a large correlation between the nonsignificant events.

4.0 Controlling The Errors: The Sequential Probability Ratio Test

Previous space-time clustering approaches (Williams 1984) require individual case data or assumptions of a Poisson distribution for time clustering or use the of correlation for spatial clustering. We have found these requirements inadequate for clustering of biosurviellance data and the aggregation typically found in these data. We have focused on the sequential probability ratio test (SPRT) (Wald 1947) for anomaly detection (Schoonewelle 1995).

For an analyst to make a decision about a possible cluster in space and time they need to understand the errors that could result in calling a cluster significant. We use the SPRT to sequentially gather and combine information over space and time to make a decision that satisfies desired error rates. Surveillance data often contains dependencies in space and time, so we discuss an extension to the SPRT for handling dependent data. This concept of sequentially gathering and combining information to make a decision to satisfy desired error rates fits well within the space-time clustering problem.

4.1. Wald's SPRT

The SPRT was originally designed to decide between two hypotheses. The test uses observations $x_1, x_2, \ldots, x_n$ to decide in favor of hypothesis H_0 or H_1. Let the variable x represent a random variable with a probability density function (PDF) $f(x \mid \theta)$, where θ represents the parameter we want to test. Here, we want to test the hypothesis H_0 that $\theta = \theta_0$ against the hypothesis H_1 that $\theta = \theta_1$. Wald's test is analogous to the Neyman Pearson's nonsequential test (Neyman 1928). It has been shown that for a fixed number of independent observations, n, the optimal test depends on the likelihood ratio Λ_n where

$$\Lambda_n = \prod_{i=1}^{n} \frac{f(x_i \mid H_1)}{f(x_i \mid H_0)} \text{ or } \Lambda_{\mathrm{n}} = \frac{f(X \mid H_1)}{f(X \mid H_0)}, \text{ where } X = [x_1, x_2, \ldots, x_n]. \qquad \textbf{(4-1)}$$

The test decides to accept or reject H_0 based on comparing Λ_n to a constant threshold. Choosing the proper threshold gives the test a desired significance level α or the probability of accepting H_1 when H_0 is true. In principle, selecting n gives the test a desired power $(1-\beta)$, where β defines the probability of accepting H_0 when H_1 is true.

Wald's test procedure has a property analogous to this optimal property except it uses two constant thresholds *A* and *B*. Here, we keep sampling as long as:

$$B < \Lambda_n < A. \qquad \textbf{(4-2)}$$

We stop taking observations and reject the hypothesis H_0 as soon as

$$\Lambda_n \geq A, \qquad \textbf{(4-3)}$$

and we stop taking observations and accept the hypothesis H_0 as soon as

$$\Lambda_n \leq B. \qquad \textbf{(4-4)}$$

The constants *A* and *B* represent the upper and lower stopping boundaries respectively. Proper selection of these boundaries produces the desired error rates α and β. In general,

$$A = \frac{(1-\beta)}{\alpha} \text{ and } \mathrm{B} = \frac{\beta}{1\text{-}\alpha}. \qquad \textbf{(4-5)}$$

The number of observations required by the sequential test to reach the desired error rates is not predetermined, but is a random variable, since the decision to stop sampling is determined by the observations made so far. We can compute the average number of necessary observations given the true hypothesis (Wald 1947). If we let

$$\lambda = \frac{f(x \mid H_1)}{f(x \mid H_0)} \text{ and } z = \log(\lambda) \qquad \textbf{(4-6)}$$

and using

$$E\{\log(\Lambda_n) \mid H_{\mathrm{k}}\} = E\{n \mid H_{\mathrm{k}}\} E\{z \mid H_{\mathrm{k}}\} \text{ for } k = 0,1 \qquad \textbf{(4-7)}$$

then the expected number of observations for deciding hypothesis H_1 is

$$E\{n \mid H_1\} = \frac{(1-\beta)\log(A) + \beta \log(B)}{E\{z \mid H_1\}} \qquad \textbf{(4-8)}$$

and the expected number of observations for deciding hypothesis H_0 is

$$E\{n \mid H_0\} = \frac{\alpha \log(A) + (1-\alpha)\log(B)}{E\{z \mid H_0\}}. \tag{4-9}$$

4.2. SPRT for Cluster Detection

In the space-time clustering problem, we have a list of adverse-health events with each item in the list containing a date and region of the event. The list of adverse-health events represent a pattern that we want to classify as background or anomalous. Here, we set θ_0 to the background class and θ_1 to the anomaly class. Thus H_0 represents the hypothesis that the pattern comes from the background and H_1 represents the hypothesis that pattern comes from an anomaly. The error rate α represents the probability of alarming on the background and calling it an anomaly or the FA rate, and the error rate β represents the probability of missing an anomaly and calling it the background or the MD rate. The observations $x_1, x_2, \ldots, x_n$ represent the number of adverse-health events in a specific region and during a specific period of time. Note this approach does not preclude the x_i observations from representing disparate observations from other data sources such as a 911 calls or over-the-counter pharmacy records. Since we could have feature vectors from other sources, the features do not necessarily have an identical distribution, and we revise the likelihood Λ_n to:

$$\Lambda_n = \prod_{i=1}^{n} \lambda_i, \text{ where } \lambda_i = \frac{f_i(x_i \mid H_1)}{f_i(x_i \mid H_0)} \tag{4-10}$$

Here $f_i(\mathbf{x}_i \mid \theta)$ represents the PDF for the observation x_i given class θ, which represents the background or anomaly class. We find it more convenient to work in the log-likelihood space:

$$Z_n = \log(\Lambda_n) = \sum_{i=1}^{n} z_i, \tag{4-11}$$

$$\text{where } z_i = \log(\lambda_i) \;. \tag{4-12}$$

We call z_i the *weight of evidence*, and if $f(x_i \mid H_1) < f(x_i \mid H_0)$ giving $z_i < 0$ then we say feature x_i leads to *negative weight of evidence* for a cluster and if $f(x_i \mid H_1) > f(x_i \mid H_0)$ giving $z_i > 0$ we say the feature leads to *positive weight of evidence* for a cluster. The test (or (4-2) thru (4-4)) then becomes:

$$\begin{array}{ll} \text{Reject } H_0 & \text{If } Z_n \geq a \\ \text{Accept } H_0 & \text{If } Z_n \leq b \\ \text{Get more data} & \text{If b} \leq Z_n \leq a \end{array} \quad , \tag{4-13}$$

where

$$a = \log(\text{A}) \text{ and b} = \log(\text{B}). \tag{4-14}$$

Figure 4-1 shows an example of a Monte-Carlo simulation of a SPRT. The independent anomaly and background features belong to two highly overlapping distributions. Let N (0,1) (Normal with mean 0 and standard deviation 1) represent the anomaly distribution and N (0,2) the background distribution. The light gray region labeled "Background" contains 99% of the background *traces*. A trace represents the log likelihood ratio output of the SPRT after testing each observation. The darker region labeled "Anomaly" contains 99% of the anomaly traces. The jagged lines represent a specific example trace of the log likelihood ratio for samples coming from an anomaly or background distribution. The two horizontal dark lines represent the stopping boundaries, labeled "a" and "b". These lines come from assuming a FA error and MD error of 0.001. When a trace goes above the *a* stopping boundary, the SPRT declares it an anomaly and when the trace goes below the b stopping boundary the SPRT declares it background. Even though the standard deviation of the test increases, the mean of the test increases faster allowing improved detection as the system receives more observations. Note the test will terminate with probability of one (Wald 1947).

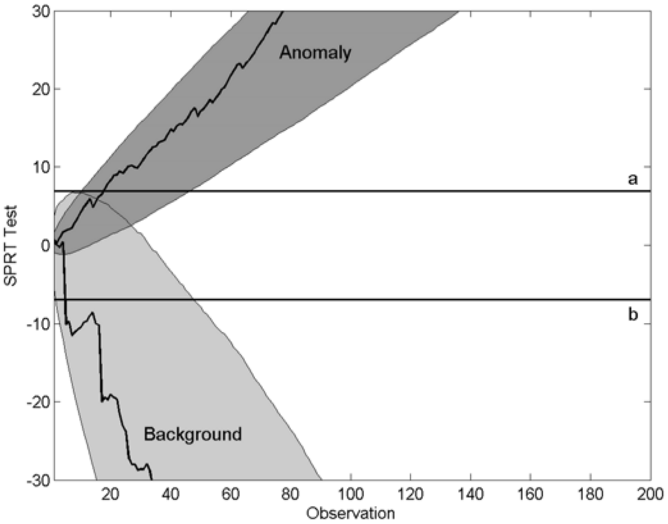

Figure 4-1. Monte-Carlo simulation of a SPRT

4.3. Feature Dependence

In the SPRT, the solution for dependent features determines the joint densities for each set of dependent features. Large numbers of dependent features can make this a challenging and a nearly intractable problem requiring inordinate amounts of historical background training data. If we incorrectly assume independent features, then we will have higher than expected error rates due to the incorrect upper and lower boundaries *a* and *b*.

For example, Figure 4-2 shows the Monte-Carlo simulation of Figure 4-1 modified to make the anomaly and background observations dependent. The Figure shows that the low value of the upper stopping boundary *a* (dotted line) causes a portion of the background traces to be labeled as anomaly. This makes the FA rate increase.

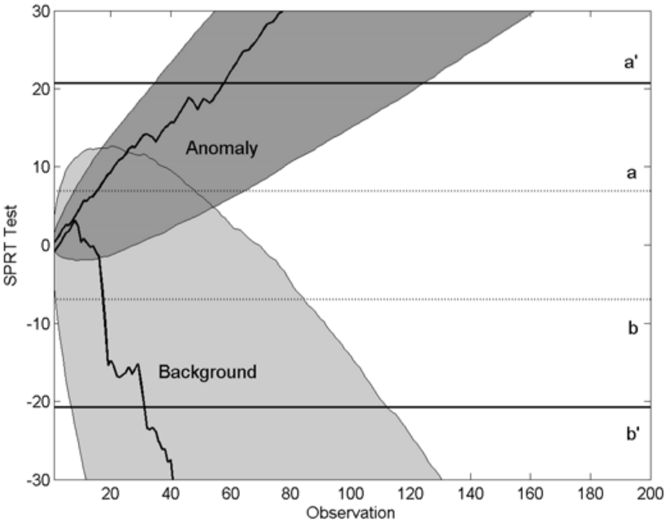

Figure 4-2. Monte-Carlo simulation of dependent observations from anomaly and background.

Our solution, to the dependent data problem, measures the effective amount of independent observations n^* (Bayley 1946). If the observations $x_1, x_2, \ldots, x_n$ are serially correlated, weakly stationary, and come from a mesokurtic (between peaked and flattened) distribution then:

$$n^* \approx \frac{n}{\kappa}, \text{ and } \kappa = \sum_{\tau=-\Delta}^{\Delta} \rho_\tau \text{ for } n >> \Delta \, . \qquad \textbf{(4-15)}$$

Note, we are interested in the worse case κ and we ignore any negative correlations. Here *n* represents the number of observations, ρ_τ the correlation coefficient of the observations taken τ lags apart:

$$\rho_\tau = \frac{E\{(x_i - E\{x\})(x_{i+\tau} - E\{x\})\}}{E\{(x - E\{x\})^2\}} \, , \qquad \textbf{(4-16)}$$

and Δ the largest lag with a non-zero correlation. Note, when no correlation exists $\rho_0 = 1$ and $\rho_\tau = 0$ for $\tau \neq 0$ giving $n^* = n$.

If the effective number of independent observations decreases by the factor $1/\kappa$, then we expect the average number of observations needed to make a decision to increase by κ . Here, we need to make more observations to make-up for the redundant dependent information. From equations (4-8) and (4-9) if we want to increase our average number of observations to handle the dependent information then the upper and lower stopping boundaries *A*' and *B*' become

$$A' = A^\kappa \text{ or } a' = \kappa\, a \qquad \textbf{(4-17)}$$

and

$$B' = B^\kappa \text{ or } b' = \kappa\, b \, . \qquad \textbf{(4-18)}$$

These adjustments in the stopping boundaries allow us to handle dependent information. For the Monte-Carlo example, Figure 4-2 shows the modified boundaries as solid horizontal lines. These modified boundaries allow the recognition of anomalies and background with dependent information.

Equation (4-15) deals with serial correlations and this translates directly to correlations in time as

$$\kappa_t = 1 + 2\sum_{k=1}^{n-1} \rho_k \, , \qquad \textbf{(4-19)}$$

where ρ_k gives the correlation $k > 0$ lags apart. For correlations in space, the following gives the equation for κ (Gilbert 1997):

$$\kappa_s = 1 + \rho_c (n_s - 1) \, , \qquad \textbf{(4-20)}$$

where n_s is the number of regions and ρ_c is the average correlation between the $n_s(n_s - 1)/2$ region pairs. For correlations across time and space Gilbert shows:

$$\kappa = \kappa_s \kappa_t \, . \qquad \textbf{(4-21)}$$

For $\rho_k = 0$ and $\rho_c = 0$ equation (4-21) reduces to $\kappa = 1$ giving $n^* = n$.

5.0 Modeling the Stochastic Process

From the previous section, we saw that we can use the SPRT to control the errors of a decision about the significance of a cluster. For the SPRT, we need to understand the stochastic process that generates the time series for background and also for the anomaly. This is captured in the PDF’s $f(y_R(t) \mid \theta_0)$ and

$f(y_R(t)\,|\,\theta_1)$. Here $y_R(t)$ represents the aggregated number of cases of an adverse health event in region R at time t, with θ_0 representing the background class and θ_1 the anomaly class.

Frequently one assumes the incidence of disease comes from a Poisson distribution, and for large populations during non-epidemic periods we would expect a Normal distribution to approximate the Poisson. However, we have found these distributions do not apply to the data sets we investigated.

5.1. French Flu Morbidity Data

For example, from http://www.b3e.jussieu.fr:80/sentiweb/en/ we have downloaded data collected by the French Sentinel Disease Network (Valleron 2002). These data contain weekly morbidity for eight diseases including influenza-like-illness. Sentinel general practitioners (about 1% of the French general practitioners) collect the data from 22 regions of France. Since 1984, the sentinel physicians can electronically connect to an information system in which they can send their morbidity reports. This gives us 16 years of flu morbidity data.

We have examined the weekly differences of flu morbidity $y_R(t) - y_R(t-1)$. Figure 5-1 shows a histogram of these differences normalized by population for all the regions in France. Along with the histogram is the best fitting Normal (dotted line) and Cauchy (solid line) PDF's. The Cauchy PDF does a far better job of approximating this histogram than does the Normal distribution.

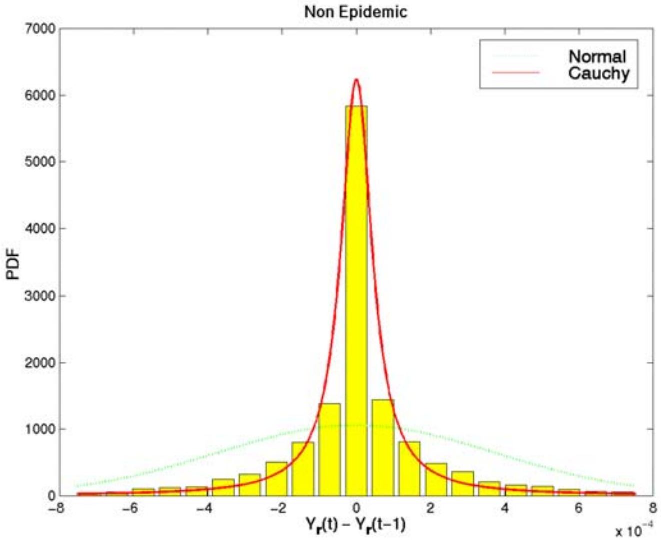

Figure 5-1. Histogram of flu morbidity differences normalized by population for all of France.

The Cauchy distribution belongs to the Levy family. Mandelbrot recognized that Levy distributions have a deep connection to scale invariant fractal random walk trajectories (Mandelbrot 1982), and Levy distributions have been used to describe many fractal processes such as search patterns of wandering albatrosses (Viswanathan 1996), human heart beat and gait (Peng 1994), stock market (Skjeltorp 2000), and diffusion processes in physics (Klafter 1996).

Since the Cauchy distribution can model a fractal process we conjecture for the French flu data set the Cauchy distribution with same parameters should generalize to different space and/or time resolutions. Indeed this appears true. Figure 5-2 shows a Cauchy distribution with the same parameters for flu morbidity differences for different populations in France (low, medium and high). Figure 5-3 illustrates the same concept with cases now aggregated over a time of two weeks.

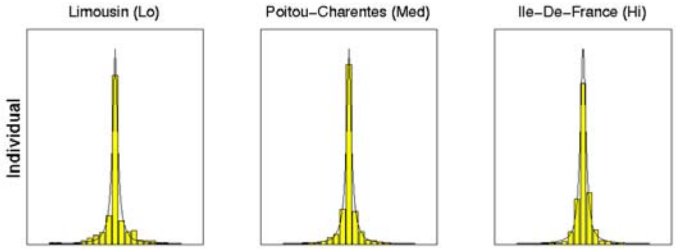

Figure 5-2. Histogram of Flu Morbidity Differences for Regions in France with Different Populations and Cauchy Distribution with the Same Parameters.

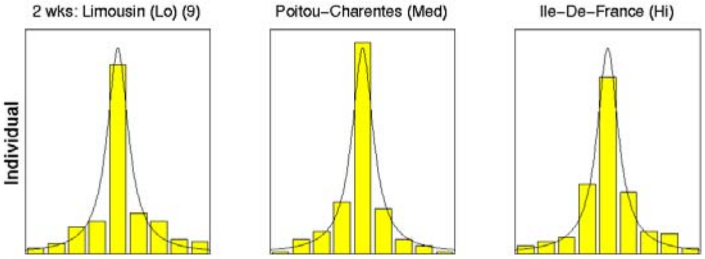

Figure 5-3. Histogram of Flu Morbidity Differences for Regions in France with Different Populations over a Two-Week Aggregation and Cauchy Distribution with the Same Parameters.

Large tails characterize Levy distributions. Figure 5-4 compares the tails of a Cauchy distribution with that of a Normal distribution in Log space. Physically, the large tails mean that we would expect small week-to-week differences of flu morbidity with occasional large random swings. These large swings become more prevalent during epidemic periods. The infinite moments of Levy distributions make it difficult to predict these swings using standard time series methods based on moments. It's important to note that we are not trying to *predict* anomalies from morbidity data, but *detect* them as early as possible. This distinction is clear in the next section when we show it's possible to find the log-likelihood ratio of the Generalized Beta Prime Distribution**.**

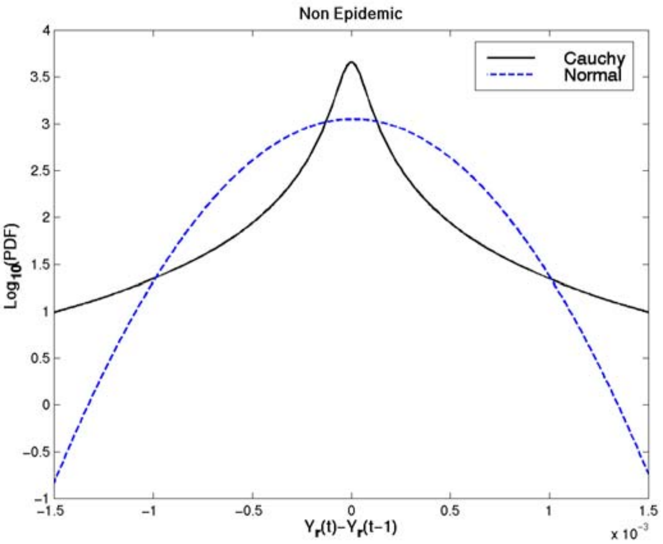

Figure 5-4. Comparison of Tails for the Cauchy and Normal Distributions.

Instead of working with differences, sometimes it's easier to work in the original morbidity space. Using the knowledge that $x_R(t) = y_R(t) - y_R(t-1)$ has the Cauchy distribution

$$f_x(x) = \frac{1}{\pi\left(1 + \frac{(x-\theta)^2}{\lambda^2}\right)\lambda} \quad \textbf{(5-1)}$$

and assuming $y_R(t)$ is stationary, we can derive the distribution for $y_R(t)$:

$$f_y(y) = \frac{4\lambda}{\pi\left(4x^2 + \lambda^2\right)}. \quad \textbf{(5-2)}$$

This distribution belongs to the family of Generalized Beta Prime (Gβ') Distributions.

Using a periodic regression model, we can determine epidemic and non-epidemic periods. Section 7.1.1 discusses this in more detail. Figure 5-5 shows a histogram of $y_R(t)$ during non-epidemic periods normalized by population with the best fitting Gβ' distribution using $\lambda = 2.90 \times 10^{-4}$. Figure 5-6 shows a similar histogram of $y_R(t)$, but for epidemic periods with $\lambda = 3.36 \times 10^{-3}$. Notice the order of magnitude increase in morbidity from Figure 5-5 to Figure 5-6 during epidemic periods. Figure **5-7** shows the Gβ' distribution for both epidemic and non-epidemics.

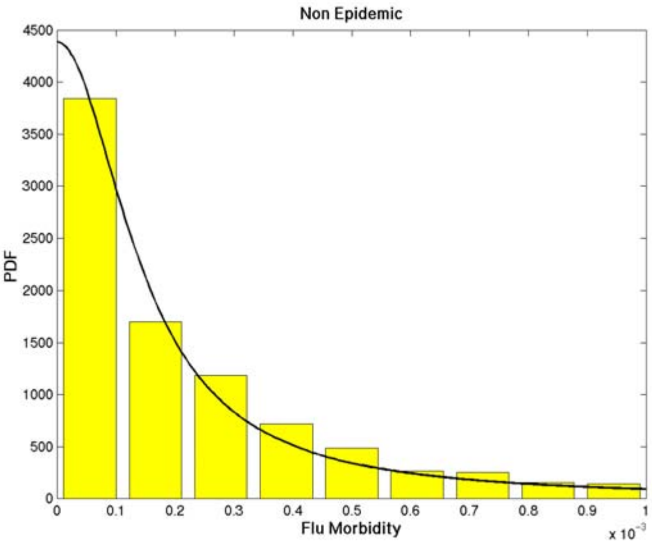

Figure 5-5. Histogram of flu morbidity during nonepidemic periods normalized by population for all of France with best fitting Generalized Beta Prime Distribution.

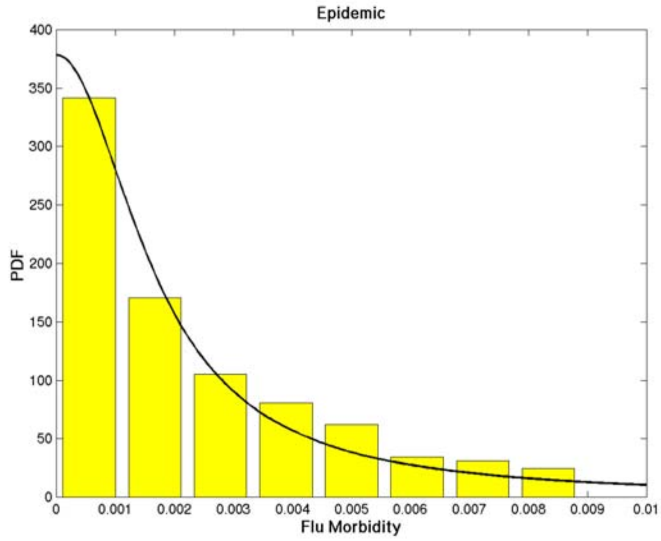

Figure 5-6. Histogram of flu morbidity during epidemic periods normalized by population for all of France with best fitting Generalized Beta Prime Distribution.

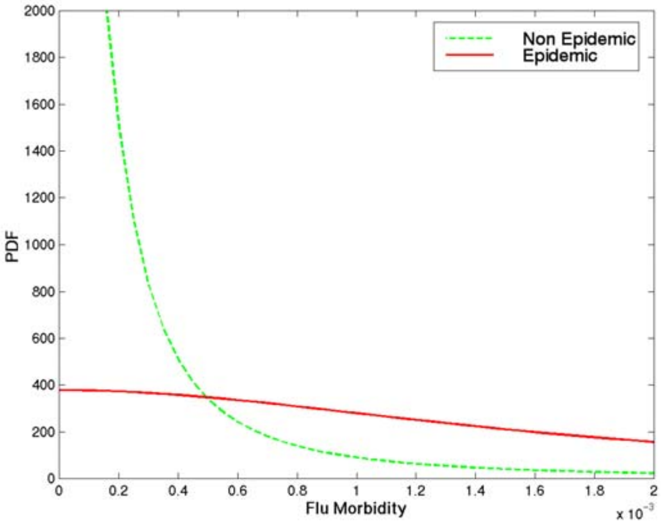

Figure 5-7. The Gβ' Distribution for French Flu during Epidemic and Nonepidemic Periods.

5.2. California Flu Morbidity Data

We also have a large data set from the California Office of Statewide Health Planning and Development (OSHPD). This database is derived from admissions to all hospitals in the state of California from 1990 to 1999. The database provides, among other information, the primary diagnosis in the form of an ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) code for each patient record along with the patient's home zipcode, the zipcode of the hospital and the date of discharge.

Figure 5-8 shows three time series of flu morbidity over 5 years from zipcodes in Los Angeles (LA) County. At most, each day and region contains one or two cases and very often no cases. The differences between the flu morbidity data from France and LA come from the fine space (zipcode vs. region level) and time (daily vs. weekly) resolutions of the OSHPD data and from the use of sentinel practitioners in France versus hospital admissions in the OSHPD data. Here, continuous distributions cannot approximate the discrete flu data from OSHPD.

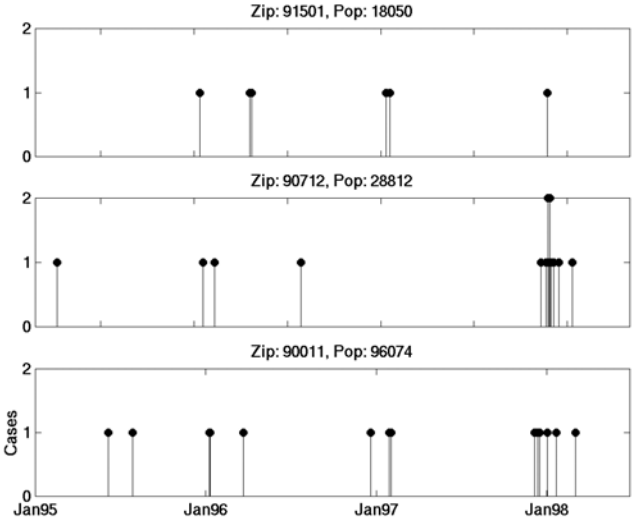

Figure 5-8. Flu morbidity from zipcodes in LA County California.

5.3. Handling Multiple Distributions: The Critical Threshold

In the previous section, we saw how different distributions arise based on how the data are collected and the space and time resolutions. While it's possible, at the very least, to derive an empirical distribution for different data sets, empirical distributions make it difficult to do the power analysis required to control the MD errors. Instead of using empirical distributions we transform the data into a binomial distribution using a *critical threshold*. The critical threshold gives the value of incidence for when the background likelihood equals the anomaly likelihood. Below the critical threshold the incidence of adverse-health events more likely belongs to the background and above the critical threshold the incidence of adverse-health events more likely belongs to the anomaly class.

Let $y_R(t)$ represent the number of adverse-health event cases in region *R* at time *t*. By identifying the critical threshold, τ_R, for every region *R* we can transform the random variable $y_R(t)$ with any distribution into a binomially distribution random variable. Here, let

$$x_R(t)=\begin{cases}1 & \text{If } y_R(t)>\tau_{\mathrm{R}} \\ 0 & \text{Otherwise}\end{cases}.$$

(5-3)

As an example, Figure 5-9 shows the corresponding log-likelihood ratio using the data in Figure **5-7**. The solid curve represents the log-likelihood ratio for the Gβ' distributions and the dotted curve the approximation based on the binomial transform. The critical threshold occurs when the background and epidemic PDF's intersect or when the log-likelihood ratio equals zero. This occurs when the normalized morbidity has a value of approximately 5×10^{-4}.

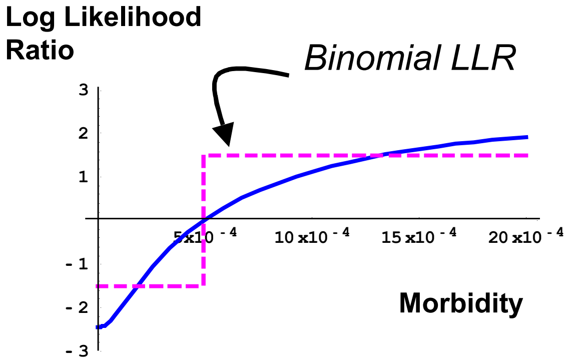

Figure 5-9. Log likelihood ratios for different levels of morbidity for the French flu data set

6.0 Binomial SPRT

Let x represent a binomial random variable that can take on the values (0,1). For example, in acceptance testing $x=0$ means a nondefective product and $x=1$ means a defective product. In our space-time clustering problem $x=0$ means the number of adverse-health events most likely belongs to the background and $x=1$ means the number of adverse-health events most likely belongs to an anomaly. Let p represent $\Pr(X=1)$. We want to determine if a sequence of x's in space and time has a p that belongs to the background or a p that belongs to an anomaly. Our solution uses a specified parameter p' and a sequential hypothesis test where the null hypothesis H_0 represents $p\le p'$ and the alternative hypothesis H_1 represents $p>p'$. These two hypotheses define a *composite* hypothesis, since the hypothesis does not completely specify the distribution. To make the hypotheses *simple*, where the distributions are completely specified, we define the tolerated risk. For the tolerated risk, we specify two values p_0 and p_1 such that $p_0<p'<p_1$. We now have two errors of a practical consequence:

1. $p<p_0$ and we reject the H_0 hypothesis

2. $p > p_1$ and we accept the H_0 hypothesis.

If $p_0 \le p \le p_1$ then we don't particularly care which decision the test makes. From p_0 and p_1 we can define our two types of error. For the type I error α we want

$$\Pr(\text{reject } H_0 | p < p_0) < \alpha \tag{6-1}$$

and for the type II error β we want

$$\Pr(\text{accept } H_0 | p > p_1) < \beta \text{ .} \tag{6-2}$$

Here, α represents the acceptable FA error and β represents the acceptable MD error. Together the quadruple $(\alpha, \beta, p_0, p_1)$ defines the tolerated risk.

We now have two simple hypotheses with H_0 representing $p = p_0$ and H_1 representing $p = p_1$. Let m represent the number of regions and times considered for a space-time cluster, x_i the result of whether the number of adverse-health events more likely belong to background or anomaly, and i an index into a specific region (zipcode, county, etc.) and time (day or week). Let the following equation define the test:

$$D_m = \sum_{i=1}^{m} x_i \text{ .} \tag{6-3}$$

Here D_m represents the number of "successes" in an experiment with m trials or we have a binomial distribution

$$\Pr(D_m = d_m) = \binom{m}{d_m} p^{d_m} (1-p)^{m-d_m} \tag{6-4}$$

where again $p = \Pr(X = 1)$. For H_0 the PDF of D_m is

$$\binom{m}{d_m} {p_0}^{d_m} (1-p_0)^{m-d_m} \tag{6-5}$$

and for H_1 the PDF of D_m is

$$\binom{m}{d_m} {p_1}^{d_m} (1-p_1)^{m-d_m} \tag{6-6}$$

For the SPRT the log-likelihood ratio becomes

$$\lambda_m = \log\left(\frac{\Pr(D_m = d_m | H_1)}{\Pr(D_m = d_m | H_0)} \right) \tag{6-7}$$

or

$$\lambda_m = d_m \log\left(\frac{p_1}{p_0} \right) + (m - d_m) \log\left(\frac{1-p_1}{1-p_0} \right) \text{.} \tag{6-8}$$

From equation (4-13) we have

$$\begin{array}{ll} \text{Reject background hypothesis} & \text{If } \lambda_m \ge a \\ \text{Accept background hypothesis} & \text{If } \lambda_m \le b \\ \text{Get more observatio ns} & \text{If b} \le \lambda_m \le a \end{array} \tag{6-9}$$

where the decision boundary thresholds a and b come from the desired error rates and are defined by equation (4-14).

From equation (6-8) we can show that

$$\lambda_{m+1} = \lambda_m + \Delta\lambda \qquad \textbf{(6-10)}$$

where

$$\Delta\lambda = \begin{cases} c, \text{if } x_{m+1} = 1 \\ d, \text{if } x_{m+1} = 0 \end{cases}, \qquad \textbf{(6-11)}$$

$$C = \frac{p_1}{p_0}, \qquad \textbf{(6-12)}$$

$$c = \log C, \qquad \textbf{(6-13)}$$

$$D = \frac{1-p_1}{1-p_0}, \qquad \textbf{(6-14)}$$

and

$$d = \log D \, . \qquad \textbf{(6-15)}$$

We call c positive evidence for an anomaly and d negative evidence for an anomaly.

Introducing an auxiliary parameter h allows us to estimate the operating characteristic (OC) function and the expected number of observations given the tolerated risk quadruple $(\alpha, \beta, p_0, p_1)$ (Wald 1947). Let p represent the true proportion of having $X = 1$, $L(p)$ the OC function or the probability of accepting H_0 for a given p, and $E(n)$ the expected number of observations required to make a decision then

$$p = \frac{1-D^h}{C^h - D^h} \qquad \textbf{(6-16)}$$

$$L(p) = \frac{A^h - 1}{A^h - B^h} \qquad \textbf{(6-17)}$$

$$E(n) = \frac{bL(p) + a[1-L(p)]}{cp + d(1-p)} \, . \qquad \textbf{(6-18)}$$

Some points on the $L(p)$ and $E(n)$ can be computed without the auxiliary parameter h :

$$L(0) = 1,\ L(1) = 0,\ L(p_0) = 1-\alpha,\ L(p_1) = \beta,\ L(s) = \frac{a}{a+|b|}, \qquad \textbf{(6-19)}$$

and

$$E(0) = \frac{b}{d},\ E(1) = \frac{a}{c},\ E(p_0) = \frac{b[1-\alpha] + a\alpha}{cp_0 + d(1-p_0)},\ E(p_1) = \frac{b\beta + a[1-\beta]}{cp_1 + d(1-p_1)},\ E(s) = \frac{ba}{cd} \qquad \textbf{(6-20)}$$

where

$$s = \frac{d}{d-c} \, . \qquad \textbf{(6-21)}$$

As an example, let the probability of $X = 1$ for the background p_0 equal 0.1 and let the probability of $X = 1$ for the anomaly p_1 equal 0.9, the missed detection error $\beta = 0.01$, and the false alarm error $\alpha = 0.01$. Figures 6-2 and 6-1 shows the resulting $E(n)$ and $L(p)$ curves with the triangles showing the points computed from equations (6-19) and (6-20). Figure 6-1 shows $E(n)$ the average number of observations for making a decision, for different values of the true proportion of $X = 1$, p. For $p < p_0 = 0.1$ or $p > p_1 = 0.9$ the expected number of observations is about 2.5 or less. The maximum number of average observations is about 4.5 at $p = 0.5$.

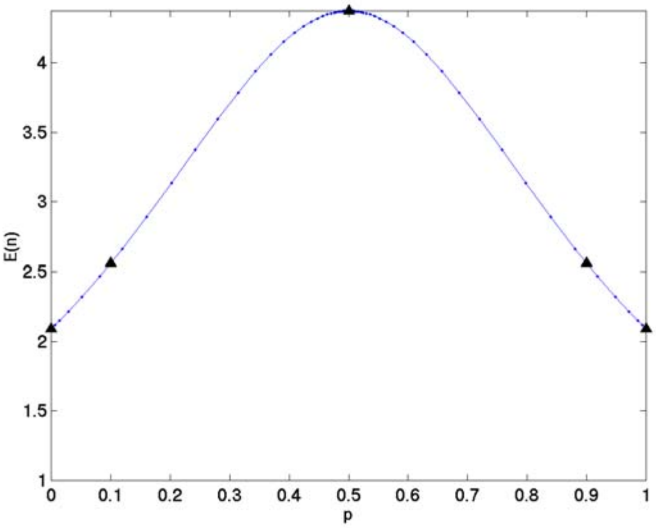

Figure 6-1. Average Number of Observations, $E(n)$, Versus the True Proportion for $X = 1$, p, for p_0=0.1, p_1=0.9, $\alpha = 0.01$, and $\beta = 0.01$

Figure 6-2 shows that for the probability of accepting H_0 (the background hypothesis) for different values of the true proportion of $X = 1$, p. For $p < p_0 = 0.1$ the probability of accepting H_0 is $1 - \alpha = 0.99$ and this probability increases as p gets smaller. This produces false alarms errors less than $\alpha = 0.01$. For $p > p_1 = 0.9$ the probability of accepting H_0 is $\beta = 0.01$ and this probability decreases as p gets bigger. This produces missed detection errors less than $\beta = 0.01$. For $0.1 = p_0 \leq p \leq p_1 = 0.9$ the test accepts H_0 with probabilities between $1 - \alpha = 0.99$ and $\beta = 0.01$. In this interval we don't care which decision the test makes. If we want to make this interval smaller then $E(n)$ would increase requiring more observations on average to make a decision. Figure 6-3 shows an example for $p_0 = 0.3$ and $p_1 = 0.7$.

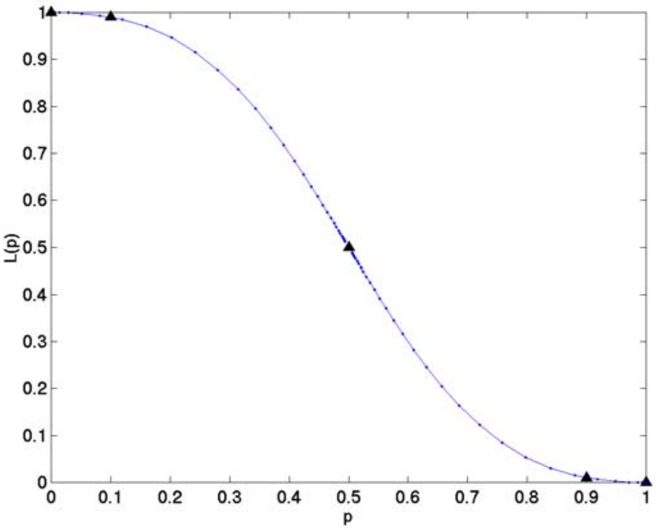

Figure 6-2. Probability of Accepting H_0, $L(p)$, Versus the True Proportion for $X=1$, p, for p_0=0.1, p_1=0.9, $\alpha=0.01$, and $\beta=0.01$

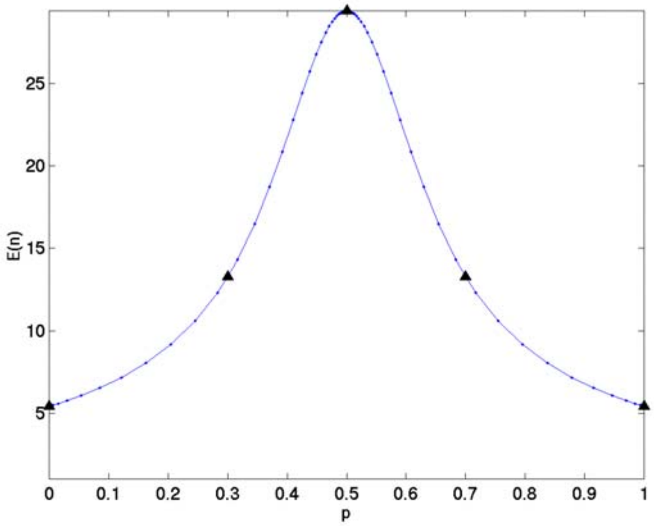

Figure 6-3. Average Number of Observations, $E(n)$, Versus the True Proportion for $X=1$, p, for p_0=0.3, p_1=0.7, $\alpha=0.01$, and $\beta=0.01$

6.1. Space-Time Clustering based on Binomial SPRT

Our approach for space-time clustering takes an analysis date and for every region searches back in time and over nearby regions for all possible space-time clusters. The user determines the maximum cluster size in time and space, and the Binomial SPRT determines the cluster's significance. Let $Y_r = [y_1, \ldots, y_m]$ represent a vector of adverse-health event cases to be considered as a possible space-time cluster centered at region *r*. Here, y_k represents the number of adverse-health event cases for a specific time and place, and k an index to a specific region (zipcode, county, etc.) and time (day or week). The algorithm searches over different sizes of clusters, *m*, with the user determining the largest *m*. The user also specifies the largest acceptable FA α and MD β error rates over all the possible cluster sequences.

To use the binomial SPRT we need to transform the vector Y_r into a binomial 0/1 vector $X_r = [x_1,\ldots,x_m]$. We accomplish this by using the critical threshold τ_k:

$$x_k = \begin{cases} 1 & \text{If } y_k \geq \tau_k \\ 0 & \text{Otherwise} \end{cases}. \qquad \textbf{(6-22)}$$

In general, τ_k varies for different regions and could depend upon, among other things, the region's population, area, average income, and doctors, hospitals, and insurance companies that service the area. For example, in determining flu epidemics high-income people might go to their doctor at the first sign of the flu more often than lower income people, or certain insurance companies might allow more tests in the diagnosis of flu. Determination of τ_k depends on the disease of interest, and the amount of historical data available, and knowledge about the anomaly. We will discuss determining τ_k more for the specific applications of space-time clustering in the section 7.0.

Next we convert X_r to evidence using equation (6-11). Conversion to evidence requires knowledge of the tolerated risk parameters p_0 and p_1. Like τ_k determination of p_0 and p_1 depends on the amount of historical data and knowledge about the anomaly. We will postpone discussion of their determination until we discuss specific applications.

To get the total evidence for the space-time cluster we add the evidence together and account for dependencies using the formulas for effective number of independent observations and equations (4-17) and (4-18). If the total evidence goes above the *a* decision boundary we call the cluster an anomaly, otherwise we don't have enough observations to make a decision or the cluster belongs to the background. The *a* decision boundary comes from the user specified false alarm and missed-detection error rates α and β:

$$a = \log\left[\frac{(1-\beta)}{\alpha}\right] . \qquad \textbf{(6-23)}$$

7.0 Results

We test our space-time clustering algorithm on two types of problems. One test is the early detection of flu epidemics, and the other is the early detection of a simulated anthrax attack.

7.1. Early detection of flu epidemics

One way to measure the performance of space-time clustering is in the early detection of flu epidemics. We test our algorithm on two flu data sets, one from France and other from CA OSHPD.

7.1.1. Determining the Start of an Epidemic

Currently, a technique developed by Serfling (Serfling 1963) determines the start of a flu epidemic in France (Toubiana 1998). The method of Serfling defines the start of an epidemic as an "*observed value of incidence above the 95% confidence threshold for a periodic-regression-model for two consecutive weeks*." This method doesn't use any space information, since the reported flu incidences are combined for the entire country. The Center for Disease Control and Prevention (CDC, 2000) also uses periodic regression models to determine the seasonal baseline for flu. Figure 7-1 shows a periodic regression model, derived using the French flu data set, to determine the start of epidemics in France. The filled circle shapes indicate the start of an epidemic. As in (Toubiana 1998) the years 1984 and 1986 have multiple starts of an epidemic and we drop those years from consideration, leaving us with 14 of the possible 16 epidemics.

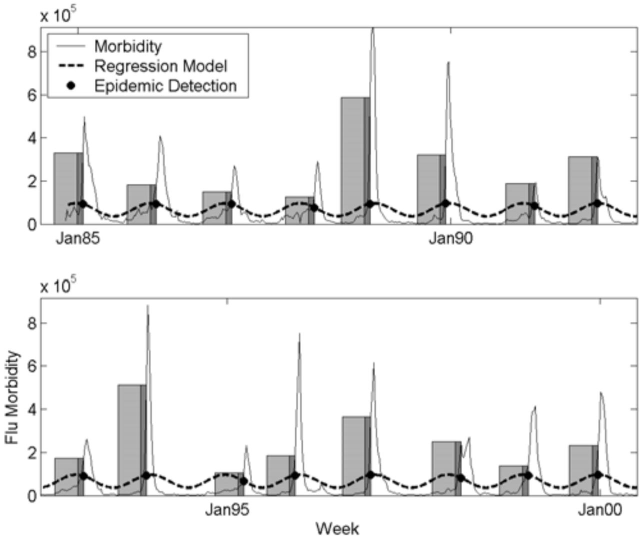

Figure 7-1. Defining the start of a flu epidemic using a periodic regression model, for the entire French data set. The darker bars represent the periods of possible correct detections and the lighter bars represent periods of possible false alarms.

Figure 7-2 shows weekly flu data from Los Angeles county California. The best-fit periodic regression model determines the start of flu epidemics, indicated by the filled circle shapes.

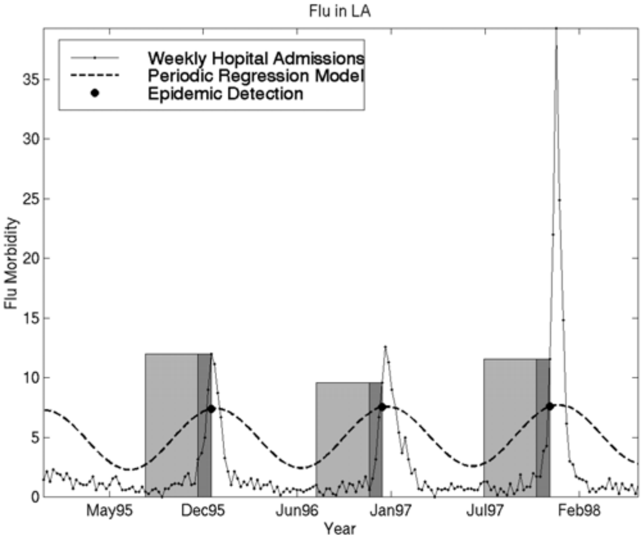

Figure 7-2. Defining the start of a flu epidemic in Los Angeles County California. The darker bars represent the periods of possible correct detections and the lighter bars represent periods of possible false alarms.

7.1.2. Measuring Performance

We assume we can't detect the start of the flu more than four weeks in advance (Toubiana 1998). The week the epidemic starts and the four weeks before define a correct detection. The darker shaded regions in Figures 7-1 and 7-2 indicate these periods. The lighter shaded regions show the FA period as any detection six months before the four-week correct detection period. Using these definitions we can determine the sensitivity (probability of detection) and specificity (1.0-<probability of false alarm>).

7.1.3. Determination of parameters

For determining algorithm parameters in the French flu data, we define epidemics as whenever the morbidity goes above the periodic regression curve plus four weeks on either side of this period. All the other weeks we define as non-epidemic. Using four years of the French flu data as training data we get the histograms and fits of the Gβ' distribution shown in Figures 5-5 and 5-6. Where these PDF's intersect gives us the critical threshold τ_{k} of 5×10^{-4} for all the regions $k = 1,2,\ldots$. Using the training data we also estimate the tolerated risk parameters p_0 and p_1. Recall for region *k* with flu morbidity y_k:

$$p_0 = \Pr(y_k > \tau_k \mid \text{Non - Epidemic}) \qquad \textbf{(7-1)}$$

and

$$p_1 = \Pr(y_k > \tau_k \mid \text{Epidemic})\,\textbf{.} \qquad \textbf{(7-2)}$$

For the CA OSHPD data, we define epidemics as whenever the weekly morbidity goes above the periodic regression curve plus 28 days on either side of this period. All the other days we define as non-epidemic. Because of the fine time resolution (daily) and space resolution (zipcode level), the number of cases is very sparse with at most 1 or 2 cases and very often zero cases. Due to the sparseness of this data set, we use a critical threshold τ_{k} of 0.5. Using two years of CA OSHPD flu data as training data, we estimate the tolerated risk parameters p_0 and p_1 for each zipcode using equations (7-1) and (7-2).

Some zipcodes don't have any cases during the non-epidemic periods. To handle this problem we estimate p_0 and p_1 using a Bayesian approach and a prior Beta PDF of

$$h(p_k) = \begin{cases} \dfrac{\Gamma(\nu_k + \omega_k)}{\Gamma(\nu_k)\Gamma(\omega_k)} p_k^{\nu_k - 1}(1 - p_k)^{\beta_k - 1}, & 0 < p_k < 1 \\ 0, & \text{Otherwise} \end{cases} \quad \text{for } k = 0,1 \qquad \textbf{(7-3)}$$

Using a posterior mean, this gives a Bayesian estimator of (Hogg 1978):

$$p_k = \frac{d_m}{m}\theta_k + \frac{\nu_k}{\nu_k + \omega_k}(1 - \theta_k) \quad \text{for } k = 0,1 \qquad \textbf{(7-4)}$$

where d_m represents the total number of days the morbidity goes above the critical threshold, *m* the number of samples, and

$$\theta_k = \frac{m}{m + \nu_k + \omega_k} \quad \text{for } k = 0,1\,. \qquad \textbf{(7-5)}$$

We estimate the constant parameters ν_k and ω_k using the method of moments and the appropriate training data from all the zipcodes. As $m \to \infty$, the parameter $\theta \to 1$ making the Bayesian estimator equal to the unbiased estimate of $\frac{d_m}{m}$. For $m = 0$ or no data, the parameter $\theta = 0$ making the Bayesian estimate equal to the mean of the Beta Distribution $\left(\dfrac{\nu_k}{\nu_k + \omega_k}\right)$.

7.1.4. Performance Results

Figure 7-3 shows a false alarm (FA) vs. probability of detection (PD) plot for detecting flu epidemics in the French flu data using space-time clustering with a binomial SPRT. We define *lead-time* as the number of weeks a significant cluster was found before the periodic regression model detection of an epidemic. The max cluster size in time is 4 weeks and the maximum cluster size in space is up to 3 adjacent regions. The

Figure shows perfect results for 0, 1, and 2 weeks of lead-time with only a very small degradation in PD for up to 3 weeks. With a lead-time of 4 weeks the PD performance drops noticeably, but is still very high.

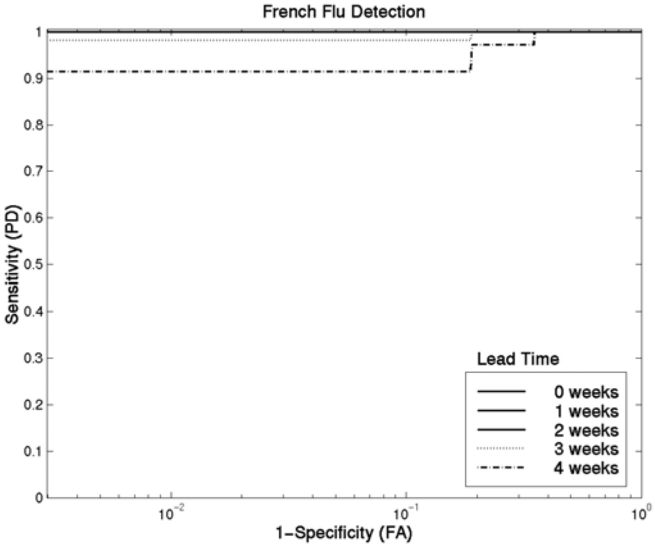

Figure 7-3. FA vs. PD for the French flu data

Figure 7-4 shows the FA vs. PD plot for detecting flu epidemics in the CA OSHPD data using our space-time clustering. The maximum cluster size in time is 21 days and the maximum size in space is up to the 50 closest zipcodes. The Figure shows perfect results for 14 days of lead-time with only a very small degradation in PD for up to 21 days. With a lead-time of 28 days the PD performance drops noticeably, but is still good.

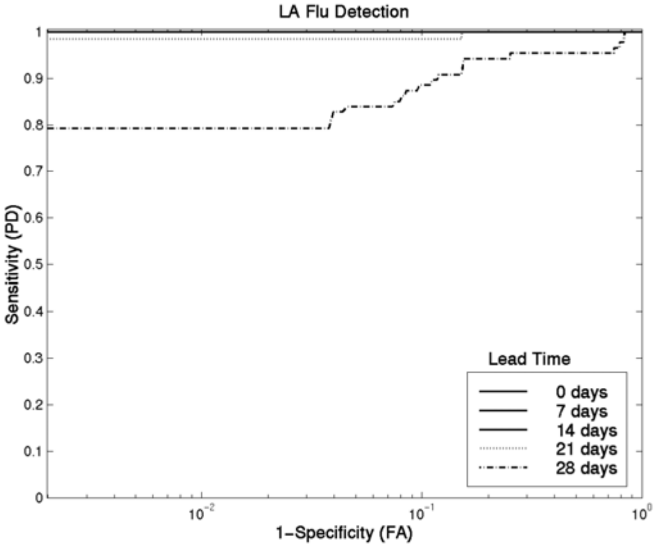

Figure 7-4. FA vs. PD for the LA flu data

7.2. Early detection of a simulated anthrax attack

For the Weapons of Mass Destruction Decision Analysis Center (WMDDAC) project, a simulation capability is being developed to enable decision makers to simulate an event and make choices over the course of the event. The final system would allow decision makers to run through several scenarios and simulate the response to their inputs. The initial focus is on biological terrorism with a prototype of a simulated anthrax attack. A simple disease model for anthrax as well as background health data (from the

OSHPD database) has been developed. A space-time clustering algorithm gives a possible tool for the decision maker. In this section, we show how space-time clustering based on the binomial SPRT could aid a decision maker in early detection of a small-scale anthrax attack.

In its early stages, a health care professional might classify anthrax as respiratory illness or major diagnostic category (MDC) 04 in the OSHPD database. Figure 7-5 shows the MDC 04 morbidity for the years 1991 through 1997 in San Francisco County and the 8 surrounding counties of Alameda, Contra Costa, Marin, Napa, San Mateo, Santa Clara, Solano, and Soma. These data provide training data and the data from the simulated anthrax attack gives a signal to add to the 1998 background data. From the Figure, one can see that the MDC 04 data has a seasonal component with peaks in the winter months and also a slight trend due mostly likely to the increasing population.

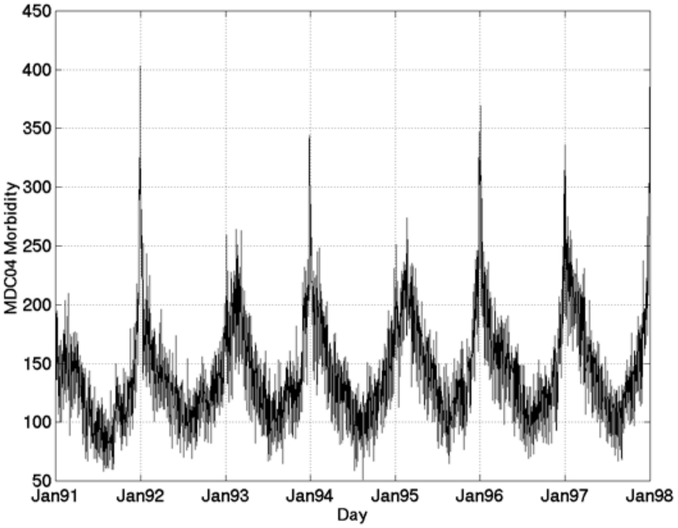

Figure 7-5. Respiratory illness or major diagnostic category (MDC) 04 from the OSHPD Database

Figure 7-6 shows the MDC 04 morbidity for four individual zipcodes. This time series appears discrete even at the higher populations. For the two smaller zipcode populations the MDC 04 incidence has at most one case with a single instance of two cases over the seven years of data. For zipcodes with larger populations some periodicity exists, not necessarily in the number of cases, but in the frequency of cases with a higher frequency occurring in the winter months.

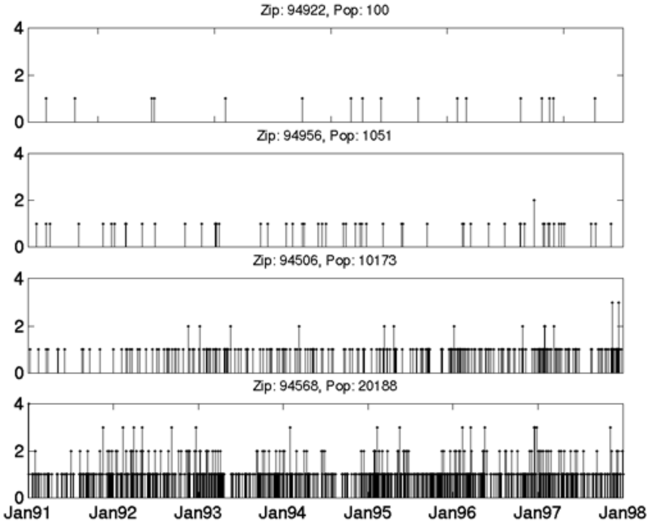

Figure 7-6. Respiratory illness MDC 04 Morbidity for four individual zipcodes in the San Francisco Bay Area, CA.

On top of this data we put the number of simulated anthrax cases computed from the WMDDAC simulator. The date assigned to a simulated anthrax case depends on the dosage a person receives and when their symptoms become severe. The transition times from exposed to mildly symptomatic to severely symptomatic to dead/recovered is variable and depends on a Monte-Carlo simulation. The zipcode assigned to the cases is determined by where the patient lives. Figure 7-7 shows 3 simulated anthrax plumes originating in the San Francisco Bay Area, CA. Shape and size of the simulated plume depend on canned weather conditions used by simulator. The two outer contours don't contain enough anthrax to cause any cases.

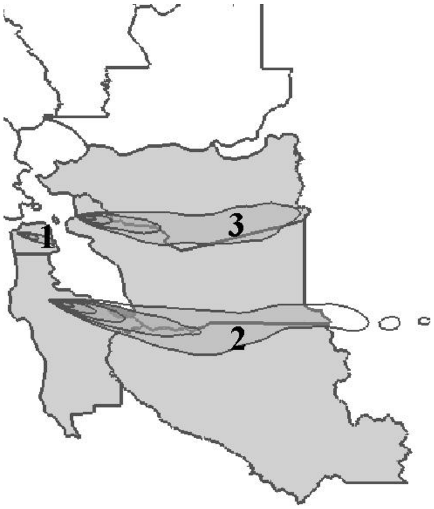

Figure 7-7. Simulated anthrax plumes originating in the San Francisco Bay Area, CA.

Figure 7-8 shows the number of anthrax cases generated by the simulated plume 2 in each of the 7 affected zipcodes. Here, only one person presents severe symptoms by day 3. Only by day 4 when 36 people become severely symptomatic do we have a chance of detecting the attack. Figure 7-9 shows the plume 2 attack starting on 1/11/1995 with the first 4 days of the attack on top of the MDC 04 data. From this aggregated data it's impossible to find the local anomaly.

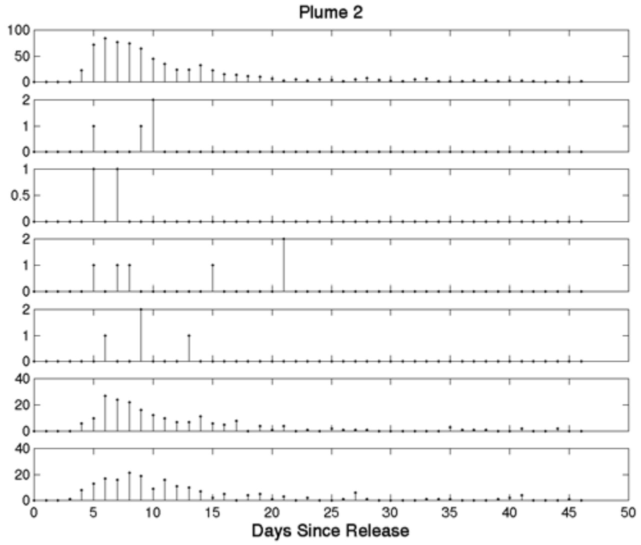

Figure 7-8. Number of simulated anthrax cases generated by plume 2.

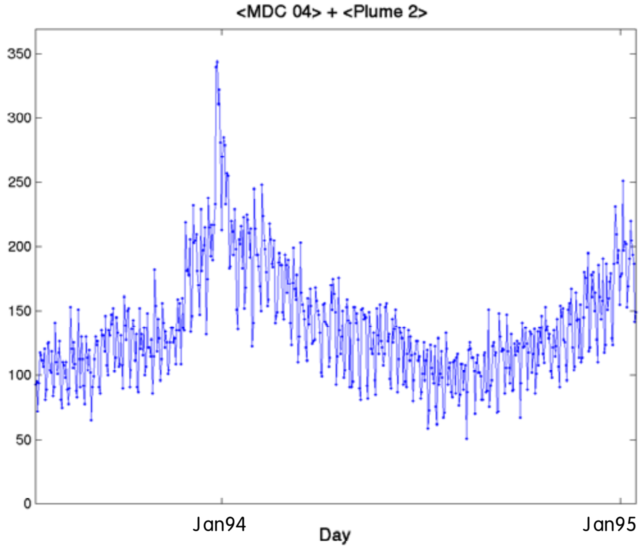

Figure 7-9. Simulated anthrax plume 2 attack starting on 1/11/1995 with the first 4 days of the attack on top of the respiratory illness MDC 04 data.

7.2.1. Modeling the Background

To find anomalies in the input data, we need to remove the seasonal and trend components from the input data. One approach uses regression techniques to fit a model to the data. Figure 7-10 shows the best-fit periodic regression model with a trend for the MDC 04 data normalized by population. While the model adequately removes the background seasonal component and trend from the data over the 9 counties, it would not work very well at the individual zipcode level where the data is more discrete and the frequency of cases marks the periodicity and not necessarily the number of cases.

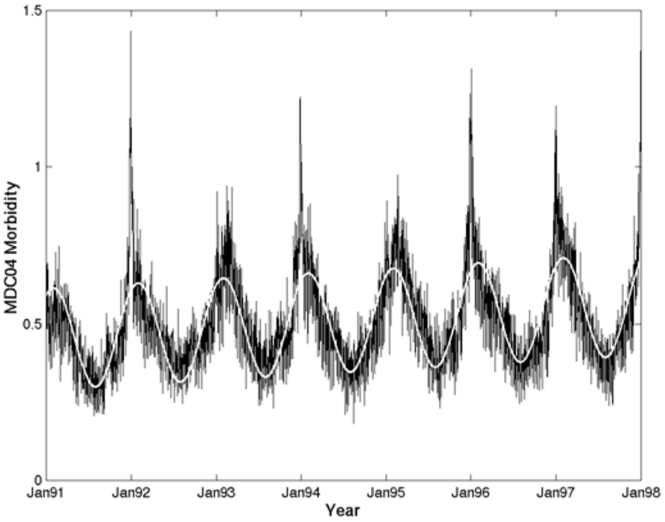

Figure 7-10. Best Fit Periodic Regression Model with a Trend for the MDC 04 Morbidity Data

We have chosen to use differencing (Box 1976) to remove the seasonal component. If $y_k(t)$ represents the number of MDC 04 cases in region k at time t, $\psi_k(t)$ the number of simulated anthrax cases, and *L* represents the periodicity of the time series then computing:

$$\delta_k(t) = \psi_k(t) + y_k(t) - y_k(t-L) \tag{7-6}$$

allows removal of the seasonal MDC 04 component. For *L*=365 days, $\delta_k(t)$ represents a new time series with the seasonal component removed. Figure 7-11 shows examples of the differenced time series for the background in 4 zipcodes of varying populations in the 9 county areas of interest. In this approach, we ignore the trend components, since it is so small from year to year.

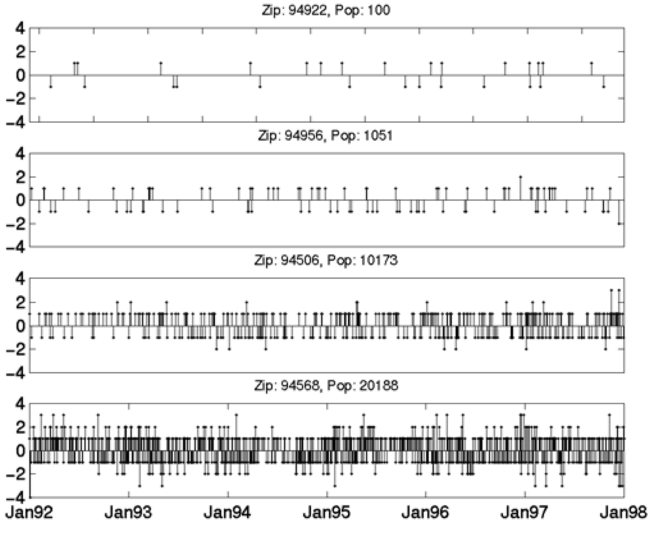

Figure 7-11. Differenced time series for the background in 4 zipcodes of varying populations in San Francisco Bay Area.

7.2.2. Determination of parameters

The binomial solution requires us to find the critical threshold τ_{k} and the tolerated risk parameters $p_0(k)$ and $p_1(k)$:

$$p_0(k) = \Pr(\delta_k > \tau_k \big| \text{MDC 04}) \qquad \textbf{(7-7)}$$

and

$$p_1(k) = \Pr(\delta_k > \tau_k \big| \text{Anomaly})\,. \qquad \textbf{(7-8)}$$

Figure 7-12 shows a cartoon of a background distribution $f(\delta \,|\, H_0)$. This distribution we can get from historical data, and we assume the distribution is symmetric about the *y*-axis. Given different critical thresholds τ we can estimate p_0. The problem is the selection of the critical threshold τ and the estimation of p_1 for all the zipcodes.

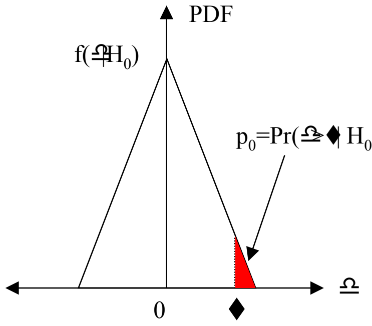

Figure 7-12. Background distribution $f(\delta \,|\, H_0)$.

To solve this problem, we introduce two new parameters c_0 and c_1 and define the zipcode with largest population as the *standard zipcode, s*. The parameter c_0 represents the number of cases above the standard zipcode's background that we are still willing to call background. The parameter c_1 defines the number of cases above the background for sounding the alarm for an anomaly. Figure 7-13 shows a cartoon illustrating the worse case background $f(\delta + c_0 \,|\, H_0)$ and anomaly $f(\delta + c_1 \,|\, H_1)$ distributions with knowledge of the two parameters c_0 and c_1.

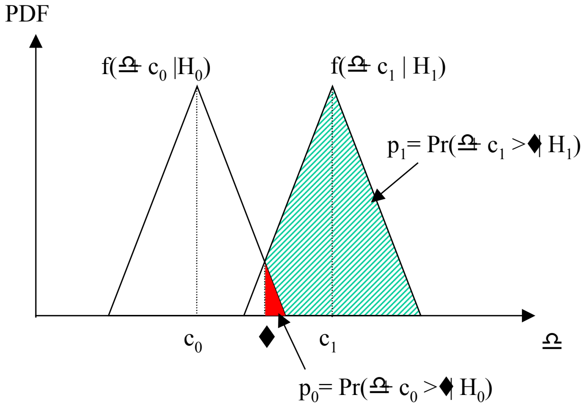

Figure 7-13. The Worse Case Background $f(\delta + c_0 \,|\, H_0)$ and Anomaly $f(\delta + c_1 \,|\, H_1)$ distributions with Knowledge of the Two Parameters c_0 and c_1.

From the Figure and assuming symmetry about the mean, the critical threshold for the standard zipcode is just the average of c_0 and c_1:

$$\tau_s = \frac{c_0 + c_1}{2}. \qquad \textbf{(7-9)}$$

To compute p_0 we translate the distributions in Figure 7-13 so that c_0 equals 0 with a new threshold or *virtual* threshold of $\tau_v = \tau_s - c_0$. The parameter p_0 can be estimated from historical data using this virtual threshold of

$$\tau_v = \frac{c_1 - c_0}{2} \qquad \textbf{(7-10)}$$

and because of symmetry $p_1 = 1 - p_0$. Based on the standard deviation (σ_k) of the other zipcodes, we scale the standard zipcode's c_0 and c_1 to get the c_0 and c_1 for other zipcodes. Having c_0 and c_1 for all the zipcodes allows us to compute their critical thresholds and probabilities.

As with the estimation of p_k and due to the sparseness of the data some zipcodes, we use a Bayesian approach (Hogg 1978) to estimate the standard deviations σ_k of δ_k for each zipcode *k*. Using the Inverted Gamma Distribution as the prior PDF:

$$g(\sigma^2) = \begin{cases} \dfrac{\lambda^c}{\Gamma(c)} \left(\dfrac{1}{\sigma^2}\right)^{1+c} e^{-\lambda/\sigma^2}, & \sigma^2 > 0 \\ 0, & \text{Otherwise} \end{cases} \qquad \textbf{(7-11)}$$

we get a Bayesian estimator of:

$$\sigma^2 = y\theta + \frac{\lambda}{c-1}(1-\theta) \qquad \textbf{(7-12)}$$

where y represents the unbiased estimate of the variance, *m* the number of samples, and

$$\theta = \frac{m}{m + 2c - 2}. \qquad \textbf{(7-13)}$$

We estimate the constant parameters λ and c using the method of moments and the appropriate training data from all the zipcodes. As $m \to \infty$, the parameter $\theta \to 1$ making the Bayesian estimator equal to the unbiased estimate. For $m = 0$, the parameter $\theta = 0$ giving the Bayesian estimate as the mean of the Inverted Gamma Distribution $\left(\dfrac{\lambda}{c-1}\right)$.

Figure 7-14 shows the result of picking different virtual thresholds (7-10), computing p_0 and $p_1 = 1 - p_0$ from the MDC 04 data of the largest zipcode *s*, and then computing the expected number of observations to a decision $E(n)$ (6-18). The graph shows that as the virtual threshold gets smaller or as c_0 and c_1 get closer it takes more observations to detect the difference between anomaly and background. Since we want to detect an anomaly as soon as possible and we don't expect to have small anomalies around long enough to detect, we advocate using a virtual threshold of 6 or 7. Small values for c_0 of 1, 2, or 3 also appear adequate for detecting an attack early without a lot of false alarms.

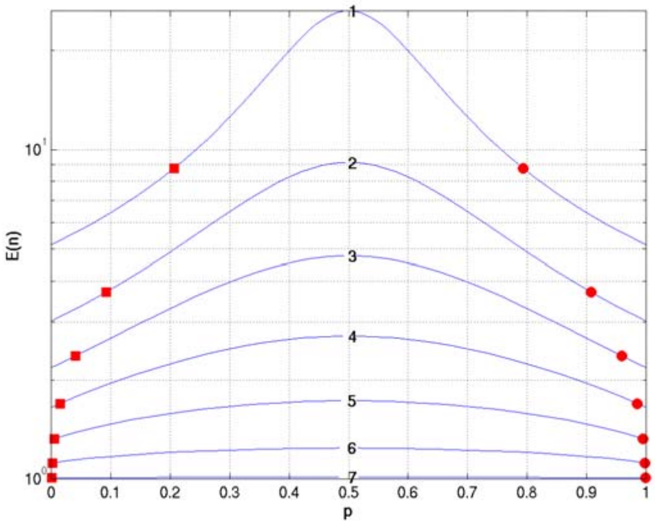

Figure 7-14. Expected Number of Observations to Make a Decision for the Standard Zipcode for Different Virtual Thresholds.

7.2.3. Results

Table 1 shows the number of simulated anthrax cases for each plume versus the days after the start of the attack. The table shows that in most cases infected people do not become severely symptomatic until at least the fourth day after the attack, and for the other cases it's the third day. Simulated plumes 4 and 10 produce very few cases due to the small attack size and its location in a low populated area. These attacks disappear into the MDC 04 background.

Table 1. Number of Simulated Anthrax Cases for each Plume Versus the Days after the Start of the Attack.

Number of Cases		Days After Simulated Anthrax Attack									
		0	1	2	3	4	5	6	7	8	9
Plume #	1	0	0	0	2	13	32	54	57	42	39
	2	0	0	0	1	36	98	129	118	118	102
	3	0	0	0	2	20	54	72	83	70	40
	4	0	0	0	1	0	2	3	0	2	2
	5	0	0	0	0	5	25	39	28	32	19
	6	0	0	0	1	3	19	14	15	10	18
	7	0	0	0	0	18	48	47	52	32	48
	8	0	0	0	0	3	9	15	24	24	9
	9	0	0	0	0	12	38	75	61	60	41
	10	0	0	0	0	0	4	2	3	4	2
	11	0	0	0	0	2	13	13	12	9	5
	12	0	0	0	0	3	7	11	12	6	6
	13	0	0	0	1	26	64	76	69	51	60

To estimate FA probabilities we look for space-time clusters in the MDC 04 background from 7/1/1992 to 6/30/1994. This gives $2 \times 365 = 730$ FA possibilities. To estimate the probability of detecting an attack we took the next three years of data and added an attack on one day. We did this for every possible day, which gave $3 \times 365 = 1090$ possible attacks for each plume.

Using c_0 =1 and c_1 =15 and the above test procedure, we compute the performance of the algorithm. Figure 7-15 shows two graphs. The top graph plots the morbidity of the background from 7/1/1992 to 6/30/1994, and the bottom graph shows the space-time clustering response to the background for each day. The response is the maximum SPRT value over all the zipcodes and possible clusters. Since we removed the seasonal variation, the background response appears fairly uniform over the 2 year time period. The maximum SPRT background response of 7.36 seems inordinately high since that would require a SPRT false alarm error rate of approximately 6.5×10^{-4} . The reason stems from the maximum operation. If U is the maximum of m independent observations with each observation having a cumulative distribution function (CDF) of $F(\cdot)$ then the CDF for U is:

$$F_U(u) = \Pr(U \le u) = F(u)^m \ . \qquad \textbf{(7-14)}$$

From equation (7-14), the relationship between the false alarm error of $\alpha_{\max}$ resulting from the maximum response over all zipcodes and possible clusters, and the underlying false error of α is

$$\alpha_{\max} = 1 - (1 - \alpha)^m \ . \qquad \textbf{(7-15)}$$

For our problem, we estimate m to have a value of at least 20. This gives an $\alpha_{\max}$ of 0.01.

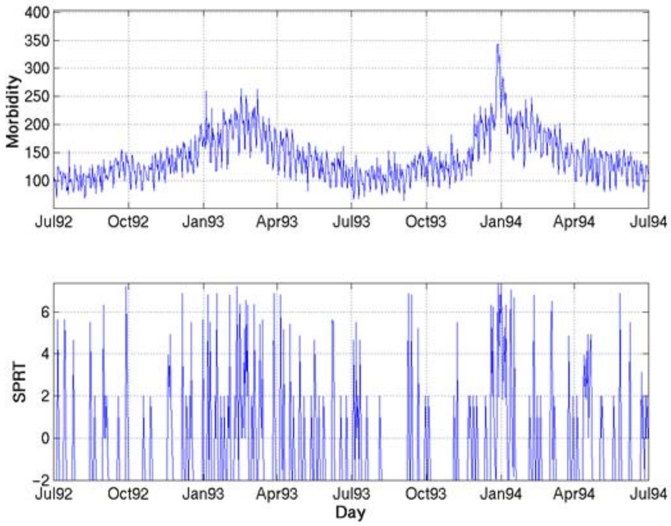

Figure 7-15. Space-Time Clustering Maximal Response over all Zipcodes and Possible Clusters.

Figure 7-16 shows the estimated actual false alarm error plotted against the theoretical error. The theoretical error easily bounds our actual error and we conclude the space-time clustering models conservatively estimate the true models.

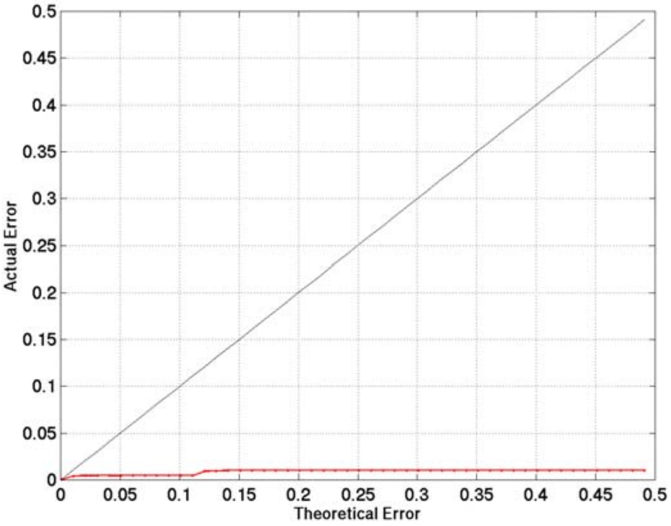

Figure 7-16. Actual Error Versus the Theoretical Error for SPRT response to the MDC 04 Background.

Referring back to Table 1, the dark gray areas indicate when the space-time clustering algorithm detects a cluster with no false alarms and greater than 99% probability of detection. The lighter gray areas indicate detection, but at a reduced PD (above 90%). The table shows we do very well at detecting an attack by the second or third day of when infected people start becoming severely symptomatic.

8.0 Conclusion

Our goal is to use on-line electronic medical information to detect and characterize a biological weapons attack after only a relatively small number of days (1-2) have passed from when the victims first begin to present symptoms. To sift through the massive amount of medical records and focus attention on potentially suspicious places and times, we use space-time clustering. Space-time clusters can indicate exposures to infectious diseases or localized exposures to toxins, and help pinpoint the location of the source of the contaminant.

We have presented statistical tests for finding space-time clusters in medical data. Current techniques for space-time clustering assume individual patient or point data, whereas most electronic medical information aggregates the space and/or time information. These space-time clustering algorithms, usually based on correlation, only indicate the presence of a cluster, but not its location and size. We have developed clustering approaches that operate on aggregated data and have embedded the test in a sequential probability ratio test (SPRT) framework. Most space-time clustering algorithms only indicate the presence of a cluster, but not its location and size. We have extended these clustering approaches to aggregated data and have embedded the test in a sequential probability ratio test (SPRT) framework. The real-time and sequential nature of health data makes the SPRT an ideal candidate. The SPRT keeps gathering and combining observations as long as the statistical test has a value between the upper stopping boundary, A, and the lower stopping boundary, B. Once the test goes above A or below B, the SPRT cluster-detector makes a decision. These upper and lower stopping boundaries determine the desired FA error rate and the desired missed-detection error rate. We have extended the SPRT to handle the spatial and temporal dependencies often found in space-time data. The resulting test not only allows us to control the error rates, but also pinpoint the cluster's location in space and time.

Most space-time clustering algorithms assume an underlying Poisson distribution. We have found that the data are far from Poisson and often best described as a Levy process. Levy distributions often characterize fractal processes and have large tails, infinite moments, and non-closed form density functions. Data collected at the zipcode and day resolution can be sparse with no more than one or two events per location and often no events. These two reasons have forced us to develop statistical tests where very little is assumed about the distribution underlying the disease process.

As a surrogate to bioterrorism data, we have experimented with two flu data sets. One set comes from the French Sentinel Disease Network and the other from the California Office of Statewide Health Planning and Development. In France, Sentinel general practitioners collect the data from 22 regions and record the week and region of the patient presented flu symptoms. The CA database is derived from admissions to all California hospitals from 1995 to 1999. For each patient record, the database provides the primary diagnosis, the admission date, and the zipcode of the patient and hospital. For both databases, we show that space-time clustering can detect a flu epidemic up to 21 to 28 days earlier than a conventional periodic regression technique.

We have also tested on simulated anthrax attack data on top of a respiratory illness diagnostic category. Results show we do very well at detecting an attack by the second or third day of when infected people start becoming severely symptomatic and start reporting to the hospital.

9.0 References

Bayley, G. V. and Hammersley, J. M., 1946. "The effective number of independent observations in an autocorrelated time series," *Supplement to the Journal of the Royal Statistical Society*, **8**, 184-197.

Center for Disease Control and Prevention, March 10, 2000. "Update: Influenza Activity – United States, 1999-2000 Season," *Morbidity and Mortality Weekly Report*, Vol. 49, No. 9.

Chen, R., 1978. "A surveillance system for congenital malformations," 73, No. 362, *Journal of the American Statistical Association*, Applications Section, pp. 323-327.

Cliff, A. D. and Ord, J. K., 1981. *Spatial Processes: Models and Applications*, Pion Limited.

Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.

Box, G. E. P., and Jenkins, G. M, 1976. *Time Series Analysis: Forecasting and Control, Revised Edition*, Holden-Day, San Francisco.

Geary, R. C., 1954. "The contiguity ratio and statistical mapping*," The Incorporated Statistician*, vol. 5, 115-145.

Gilbert, R. O., 1987. *Statistical Methods for Environmental Pollution Monitoring*, John Wiley and Sons.

Grimson, R. C., Wang, K. C. and Johnson, P. W. C., 1981. "Searching for hierarchical clusters of disease: spatial pattern of sudden infant death syndrome, *Soc. Sci. Med.*, vol. 15D, pp. 287-291.

Hungerford, L. L., 1991. "Use of spatial statistics to identify and test significance in geographic disease patterns," *Preventive Veterinary Medicine*, vol. 11, pp. 237-242.

Hogg, R. V. and Craig, A. T., 1978. *Introduction to Mathematical Statistics*, Macmillan Publishing Co., Inc.

Jacquez, G. M., 1996. "A k nearest neighbor test for space-time interaction," *Statistics in Medicine*, vol. 15, pp. 1935-1949.

Klafter, J., Shlesinger, M. F., and Zumofen, G., 1996. “Beyond Brownian motion,” *Physics today*, *49 No. 2, pp. 33-39.*

Knox, G., 1964a. "Epidemiology of childhood leukemia in Northumberland and Durham," *British Journal of Preventive and Social Medicine*, vol. 17, pp. 17-24.

Knox, G., 1964b. "The detection of space-time interactions," *Applied Statistics*, vol. 13, pp. 25-29.

Knox, G. and Lancashire, R., 1982. "The detection of minimal epidemics," *Statistics in Medicine*, vol. 1, pp. 183-189.

Kolavic, S. A., Kimura, A., Simons, S. L., Slutsker, L., and Barth, S., August 6, 1997. “An outbreak of dysenteriae type 2 among laboratory workers due to intentional food contamination,” *JAMA-Journal of the American Medical Association, 278, No. 5, pp. 396-398.*

Lai, D., 1997. "Spatial statistical analysis of Chinese cancer mortality: a comparison study of the *D* statistic," *Scand. J. Soc. Med.*, 24, No. 4, pp. 258-265.

Mantel, N., 1967. "The detection disease clustering and a generalized regression approach," *Cancer Research*, 27, No. 2, pp. 209-220.

Mandelbrot, B., 1982, *The Fractal Geometry of Nature*, Freeman San Francisco.

McKenna S. A., Koch M., and Bilisoly, R. L., 2002. “Comparing techniques for detection of epidemics in public health surveillance data,” *Conference on Unified Science & Technology for Reducing Biological Threats and Counter Terrorism,* The University of New Mexico, Albuquerque NM.

Moran, P., 1948. "The interpretation of statistical maps," *Journal of Royal Statistical Society, series B*, vol. 10, pp. 243-251.

Moran, P., 1950. "Notes on continuous stochastic phenomena," *Biometrik*, vol. 37, pp. 17-23.

Murphy, K., and Myors, B., 1998. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, Lawrence Erlbaum Associates.

Naus, J. I., 1965. "The distribution of the size of the maximum cluster of points on a line," *American Statistical Association Journal,* pp. 532-538.

Naus, J. I., 1966. "Some probabilities, expectations and variances for the size of largest clusters and smallest intervals," *American Statistical Association Journal,* pp. 1191-1199.

Neyman, J. and Pearson, E. S., 1928, "On the use and interpretation of certain test criteria for purpose of statistical inference," *Biometrika* Pt. I, 175-240, Pt. II, 263-294.

Peng, C. K., Hausdorff, J. M., Mietus, J.E., Havlin, S., Stanley, H.E., and Goldberger, A.L., 27-30 June 1994. "Fractals in physiological control: from heart beat to gait," *Conference: Levy Flights and Related Topics in Physics. Proceedings of the International Workshop*, Nice, France.

Schoonewelle, H, van der Hagen, T. H. J. J., Hoogenboon J. E., 1995. "Theoretical and numerical investigations into the SPRT method for anomaly detection, *Annals of Nuclear Energy, 22 No. 11, pp. 731-742.*

Serfling, R. E., 1963. "Method for current statistical analysis of excess pneumonia-influenza deaths," *Public Health Report*, 78, pp. 494-506.

Skjeltorp, J.A., 2000. "Scaling in the Norwegian stock market," *Physica A, 283, No. 3-4, p.486-528.*

Stroup, D. F., Williamson, G. D., and Herndon, J. L., 1989. "Detection of aberrations in the occurrence of notifiable diseases surveillance data," *Statistics in Medicine*, vol. 8, pp. 323-329.

Tango, T., 1984. "The detection of disease clustering in time," *Biometrics*, vol. 40, pp. 15-26.

Torok, T. J., Tauxe, R. V., Wise, R. P. et. al., 1997. "A large community outbreak of salmonellosis caused by intentional contamination of salad bars," *JAMA-Journal of the American Medical Association, 278, No. 5, pp. 389-395.*

Toubiana, L. and Flahault, A., 1998. "A space-time criteria for early detection of epidemics of influenza-like-illness," *European Journal of Epidemiology*, vol. 14, pp. 465-470.

Valleron, A. J., French Sentinel Disease Network: http://www.b3e.jussieu.fr:80/sentiweb/en/, 2002.

Viswanathan G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., and Stanley, H.E., 1996, "Levy flight search patterns of wandering albatrosses," *Nature*, 381, No. 6581, pp. 413-415.

Wald, A., 1947, *Sequential Analysis*, John Wiley & Sons Inc.

Wallenstein, S., 1980. "A Test for detection of clustering over time," *American Journal of Epidemiology*, 111, No. 3, pp. 367-372.

Walter, S. D., 1994. "A simple test for spatial patterns in regional health data," *Statistics in Medicine*, vol. 13, pp. 1037-1044.

Weinstock, M. A., 1981. "A generalized scan statistic for the detection of clusters," *International Journal of Epidemiology*, 10, No. 3, pp. 289-293.

Williams, G. W., 1984. "Time-space clustering of disease," *in Statistical Methods for Cancer Studies*, R. G. Cornell, ed., Marcel Dekker Inc., pp. 167-227.

Distribution:

1	MS 0844	Wallace J. Bow, 15352
12	MS 0844	Mark W. Koch, 15352
1	MS 0735	Sean A. McKenna, 6511
1	MS 0735	Roger L. Bilisoly, 6511
1	MS 1219	Annette L. Sobel, 05907
1	MS 9201	Dawn E. Kataoka, 08114
1	MS 1207	Micheal W. Trahan, 05914
1	MS 1363	Alan P. Zelicoff, 05320
1	MS 1374	Greg A. Mann, 05327
1	MS 0188	LDRD Office (Donna L. Chavez), 1030
1	MS 9018	Central Technical Files, 8945-1
2	MS 0899	Technical Library, 9616
1	MS 0612	Review & Approval Desk, 9612 For DOE/OSTI