Wavelet-Based Monitoring for Disease Outbreaks and Bioterrorism: Methods and Challenges

Bernard L. Dillard
Department of Science and Mathematics
Fashion Institute of Technology
New York, NY 10001
dillard@fitnyc.edu

Galit Shmueli
Department of Decision, Operations & Information Technologies
and The Center for Health Information and Decision Systems
Robert H. Smith School of Business
University of Maryland
College Park, MD 20742
gshmueli@rhsmith.umd.edu

1. INTRODUCTION

Modern syndromic data are unique in that they harbor several important characteristics. Of primary significance is the fact that they are available at a *very frequent* rate, such as daily and even hourly. This is due to real-time collection (e.g. – via UPC codes), electronic reporting, and transfer. Such data also tend to be *seasonal*. For example, sales of cough medication tend to fluctuate between seasons. Winter sales are different than those in the summer, and peak sales occur around holidays. In the case of emergency room visits, arrival patterns vary according to the season. Related to seasonality is the issue of *correlation to irrelevant variables*. This suggests that some variables have some dependency on other, less interesting variables. For example, when stores close, overall sales (including grocery sales) decrease. Since people may stock up on certain items on weekends, total sales (including grocery sales) tend to increase. Finally, different series even within the same data source can vary greatly in their structure, thereby requiring a very flexible monitoring system.

Classical methods for monitoring single and multiple processes for detecting abnormal behavior have been around for at least half of a century. Historically, public health facilities relied on the monitoring of traditional data streams in their effort to gain useful information as it related to the presence of diseases. These types of traditional data in which the outbreak was confirmed were in the form of death rates, laboratory results, and emergency room diagnoses. The use of these traditional surveillance systems became limited greatly by delays in getting and analyzing the data and by delays from the waiting period needed for reports to be confirmed by testing [1]. Rather than focus primarily on these types of data, it became more advantageous to rely on using data that would house the same information as traditional data streams but would possibly detect the footprint of the disease outbreak earlier. This, then, is what we refer to as syndromic data.

Monitoring these non-traditional data streams has proven useful in the quest to detect early signs of a disease outbreak. Several examples of this syndromic data include pharmacy medication sales, hits on the world wide web of medical websites (like WebMD), grocery sales of over-the-counter (OTC) medications and other health-related products, 911 calls, nurse hotlines, and chief complaints [2]. Many researchers have agreed that monitoring the sales of OTC grocery sales can indirectly provide invaluable information concerning the tracking of certain bioagents like anthrax used in biological warfare [3]. The logic behind such a revelation is twofold: 1) Non-traditional data sources, such as grocery sales and OTC medication,

contain an early footprint of an anthrax-based attack or disease outbreak; and, 2) People who feel sick are more apt to seek self-treatment before approaching the medical system [2].

These syndromic data are also *non-stationary* and *noisy*. These characteristics make wavelet-based statistical monitoring methods especially ideal. The non-stationarity of the data suggests that the behavior of the series changes over time. The mean, variance, and auto-covariance of the series do not remain constant throughout the observed time window [2]. The noisiness suggests that some degree of smoothing is required. Wavelet methods are also advantageous when the nature of the anomaly is unknown. In many syndromic data, it is unknown how a disease outbreak will manifest itself. For instance, how would an anthrax attack affect sales of cough medication? Or how would a smallpox outbreak influence the arrival patterns of ER visits?

In Section 2, we offer a brief history of wavelets, explain why they are more advantageous, and introduce the discrete wavelet transform. Section 3 introduces a set of syndromic data and applies the univariate wavelet-based technique called Multiscale Statistical Process Control, while Section 4 discusses and illustrates the wavelet-based multivariate extension or Multiscale Principal Components Analysis. Within Sections 3 and 4, we address the issue of multiple testing in order to guard against an increased false alarm rate. Section 5 highlights major results, while Section 6 offers final discussion on the proposed methods and explores a few challenges that remain.

2. WAVELETS AT WAR

The most well-known and widely used monitoring tools in epidemiology as well as other areas have been Shewhart, cumulative sum (CuSum), and exponentially weighted moving average (EWMA) control charts. Some have been generalized to multiple processes or adjusted for serial correlation. Most of them, however, assume stationarity of the series and are efficient in detecting abnormalities only of a certain nature. The advent of wavelets has been able to address these limitations and overcome monitoring hindrances experienced by many of the mentioned classical monitoring schemes. Herein, we discuss new wavelet-based monitoring techniques that are more flexible and make less structural assumptions.

2.1 A Brief History of Wavelets

Historically, wavelets have been touted as the quintessential mathematical tool for image compression. In computer science circles, they have been lauded for their ability to flexibly adapt to shapes and patterns of the original image and re-construct them using minimal space. Through a tag-team effort of using high- and low-pass filters, wavelets produce snapshots of images while minimizing pixilated space. Because wavelets possess such a great ability of stretching and shrinking, they enjoy confronting the task of duplicating complex pictures. Over the last decade or so, the utility of wavelets has widened from this well-known idea of image compression to the relatively new area of anomaly detection. Even though several scholars have suggested that these mathematical tools may provide promising results for such detection, hardly any literature exists in which these methods are examined alongside age-old, more traditional approaches for detecting out-of-control processes. Simply said, since wavelets possess this uncanny ability to adapt and flex, they become ideal “spies” on the hunt for unknown aberrant behavior in a time series. Of course, classical approaches to modeling techniques for detecting anomalies center on autoregressive moving-average (ARMA) models and Fourier analysis.

Few strides have been made to employ wavelets in the use of monitoring for anomalous detection in health care or biosurveillance. Even though authors have suggested that these mathematical tools may provide promising results for such monitoring, hardly any literature exists in which these methods are examined alongside age-old, more traditional approaches for detecting out-of-control processes. For example, Goldenberg and Zhang use wavelets to forecast and to de-noise data, respectively [3, 4]. Shmueli seeks to address some of the challenges related to using wavelet transforms and their use in detecting outbreaks, giving special attention to issues associated with multiple testing and reducing false alarm rates [5].

Shmueli admits, however, that there remain some untapped methods for wavelet-based monitoring in other fields which remain to be explored in biosurveillance [5]. One of the fields she alludes to is chemical engineering, in which Aradyhe develops wavelet-based monitoring for fields other than bioterrorism [6]. We investigate these methods for monitoring for biosurveillance and disease outbreak. It is against this backdrop of meager research and the modern-day seriousness of bioterrorism and disease outbreak (including the H1N1 virus) that we apply these newer methods to syndromic data.

2.2 Why Wavelets Work

Wavelets are mathematical functions that break data down into different frequency components and that analyze each of the frequency components with a scale-matched resolution [7]. They are localized functions, continuous in time, that drop to zero instead of oscillating forever [8]. Since wavelets are able to decorrelate autocorrelated data, the wavelet decomposition method yields an elegant and more suitable way of representing and monitoring biosurveillance data over traditional monitoring techniques [9]. Wavelet methods are highly useful in practice because data from most of the processes are multiscale in nature due to events occurring at different locations and with different localization in time and frequency; stochastic processes whose energy or power spectrum changes with time and/or frequency, and; variables measured at different sampling rates [10]. Since our goal is to monitor the series over time in order to rapidly detect an outbreak for which the anomaly pattern is unknown, it is most important to know the timings of the different frequencies. Wavelets have the capability of identifying information regarding frequencies in the data, while capturing the information regarding when notable phenomena transpire.

A key reason for choosing wavelet analysis over other classical techniques is that wavelets overcome the stationarity assumption that is the backbone of methods such as autoregressive moving-average (ARIMA) models. Since the data is frequently recorded, it becomes preferable for the purpose of rapid detection; however, data that is recorded too frequently are too noisy and requires more attention [2]. Essentially, wavelet analysis becomes more suitable than these traditional methods for several reasons:

1. Wavelet analysis allows us to analyze a series while simultaneously preserving temporal and spatial information; other key methods either preserve temporal or spatial information, not both.
2. Wavelet analysis is more flexible in its monitoring of frequent data (i.e. – data that is daily, hourly, etc.)
3. Wavelet analysis requires the least tweaking from the non-statistician user; current software makes for a user-friendly environment to aid in the use of wavelet analysis.

2.3 The One-Dimensional Discrete Wavelet Transform

Modeling frequent data therefore warrant a very flexible method such as using the one-dimensional discrete wavelet transform (DWT). By definition, the DWT is described by equation (1):

$$W_x(a,b) = \frac{1}{\sqrt{a}} \sum_{t=-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right), \quad (1)$$

where $W_x(a,b)$ is the set of transformed wavelet coefficients at the appropriate approximation and detail levels, while $x(t)$ represents the data points of the original series at time t. We express $\psi(t)$ as the basic wavelet function onto which the signal is analyzed. When stretched or shrunk, $\psi(t)$ becomes $\psi(\frac{t-b}{a})$, where "a" is the dilation parameter and "b" is the translation parameter.

In this article, the basic wavelet is stretched and shrunk to the point in which it has become the Haar wavelet. The DWT serves to decompose the original series into a linear combination of detail coefficients (*d1* – *dL*) and the finest approximation coefficients (*aL*), where *L* represents the number of decomposition levels used in the transform. An *L*-level DWT requires at least 2^L data points. Decomposition beyond this level creates "holes" in the analysis and sacrifices accuracy of the results. The DWT is very useful in analyzing data in such multiscale form and better captures the features of such data with a much lower need for teasing. Finally, it is preferable to employ an automated system, which requires the least tweaking from the non-statistician user, and in that respect, the DWT is preferable to Fourier and ARIMA methods.

2.4 The Decomposition

The method begins with one series, which is decomposed into multiple uncorrelated scales using a wavelet function. This wavelet function turns the information of a signal into a set of uncorrelated coefficients at multiple scales, which are used to reconstruct the original signal [12]. In addition to choosing which wavelet is appropriate for the analysis, the researcher must determine how many scales (*L*) of analysis are suitable. The series is then decomposed into two sets of *L* scales. The first contains the approximation coefficients, which capture low frequencies and crude trends of the original data series. The second contains the detail coefficients, which correspond to high frequencies and capture more detailed information. Figure 1 depicts a general wavelet decomposition tree, showing the relations between the original signal, the wavelet approximations, and the details, spanning across decomposition levels.

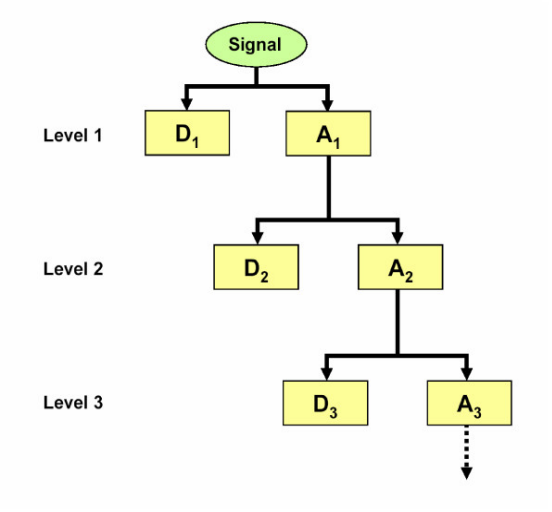

Figure 1: *Wavelet Decomposition Tree*. At each time point, the sum of the detail coefficients and last approximation level ($D_1 + D_2 + D_3 + A_3$) equals the original signal (for $L = 3$).

From a methodological point of view, DWT techniques offer an analysis of the series as a sum of orthogonal signals corresponding to different time scales. From a more practical viewpoint, from the multitude of wavelet functions that exist, our choice of wavelet in the DWT is based on the interplay between a specific analysis goal (signal processing, monitoring, etc) and the properties needed in a wavelet filter to achieve that goal [8].

Figure 2 illustrates a five-level DWT of a data series of pediatric gastrointestinal (GI) free-text chief complaints (cc) from April 10, 1998 through May 31, 2001. These complaints are a type of data collected during emergency room visits for which an admissions clerk records a patient's status on arrival to the facility; since the data entry is automated, they become amenable to typical biosurveillance systems [11]. The data shows standardized daily counts of complaints in four counties in Utah, in which 80% of the state's population lives [11]. In this DWT, we include five levels of decomposition (*L=5*). So detail coefficients (noted as *d1* – *d5*), and the finest approximation coefficients (noted as *a5*) add to produce the original series at each particular time point. In this case, level *a5* captures the overall trend of the original series, producing the general up-and-down pattern over time that exists in the data. In levels *d1*-*d3*, the wavelets do well in capturing the peak (standardized) GI cc count (denoted by *A*), which has a value of 4.35 and occurs on January 7, 2001. Essentially, this high peak manifests itself at multiple scales,

which highlights the wavelet's strength of detecting pertinent information at varying decomposition levels.

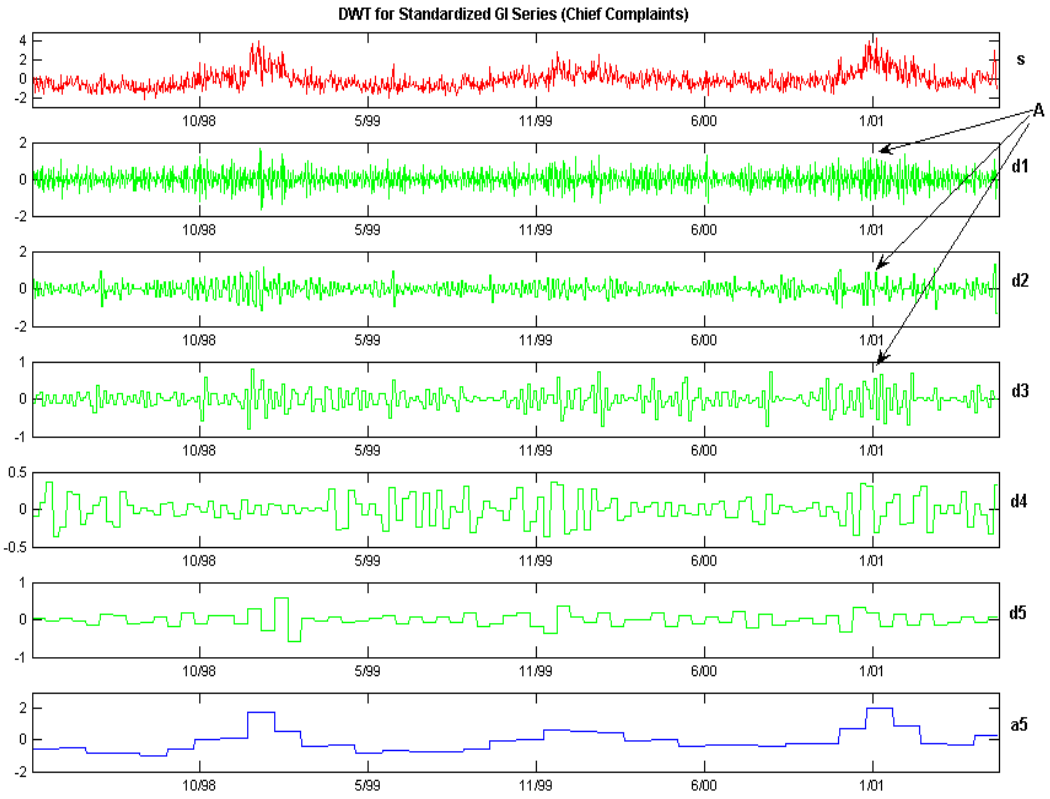

Figure 2: *DWT of pediatric gastrointestinal (GI) free-text chief complaints series (Haar wavelet)*. The GI series is denoted by *s*, and the *d*'s and last approximation level signify detail and approximation scales, respectively; here, $L = 5$.

3. MULTISCALE STATISTICAL PROCESS CONTROL

In this section, we present a detection algorithm for univariate time series that was made popular by Bakshi. The method, multiscale statistical process control (MSSPC), combines DWT and Shewhart control charts and was developed in the field of chemical engineering as work related to aberration detection [6, 10]. This method also becomes advantageous as a monitoring technique for biosurveillance or disease outbreak because the syndromic data analyzed possess the same characteristics as those stated in Section 1 (noisy, non-stationary, etc). This monitoring approach seeks to detect abnormal events at multiple scales of the series as relatively large

coefficients; the idea is to decompose the signal using DWT and then to monitor the coefficients at each scale separately using a Shewhart chart. The original series is then reconstructed from all the coefficients that exceed the thresholds at the different scales, and another Shewhart chart is used to monitor this reconstructed series. This process is illustrated in Figure 3.

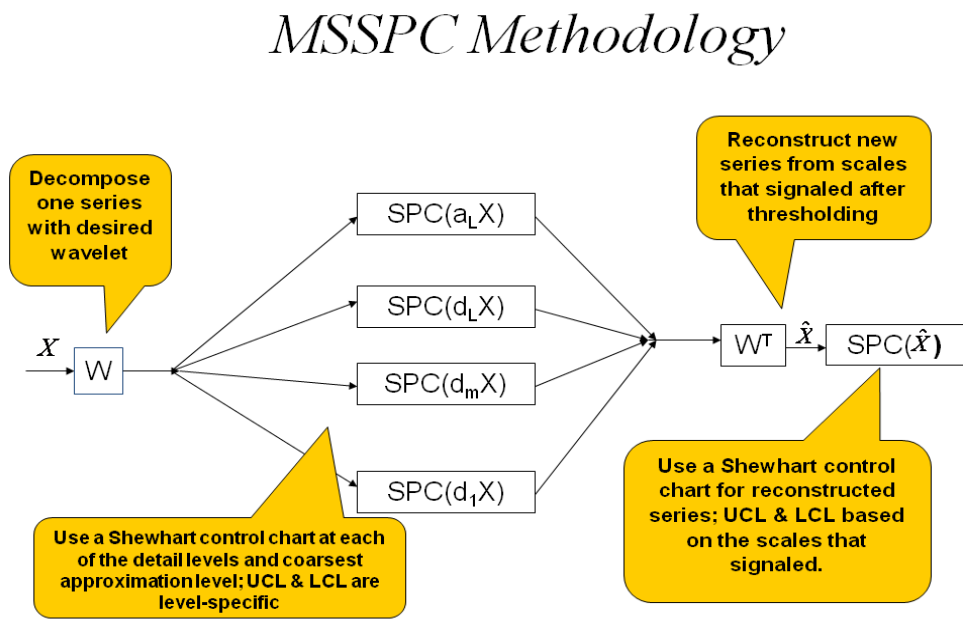

Figure 3: *MSSPC Methodology*

3.1 Scale-Specific Monitoring

Shewhart control charts are used for each level separately in order to detect abnormal coefficients relative to their scale. The center line and control limits are computed from the scale's expected value (μ_i) and its standard deviation (σ_i), where i is the scale. For the detail coefficients, $\mu_i = 0$, whereas for the approximation, μ_i is equal to the mean of the series. The σ_i's are estimated from the coefficients at scale i. The upper control limit (UCL) and lower control limit (LCL) are then $\mu_i \pm 3\hat{\sigma}_i$. This ensures that we maintain a 0.27% false alarm rate at each scale. Applying separate control charts at each scale permits identification and selection of the scales or frequency bands that appear to contain abnormal behavior [6]. The UCLs and LCLs of the Shewhart chart at each detail level and finest approximation are key in

that any coefficients that exceed these limits become potential points of alarm in our monitoring scheme.

3.2 Series Reconstruction and Alarm Detection

The next step is to reconstruct the series based on the coefficients at each scale which lie outside of the detection limits. In the reconstruction, all detail and approximation coefficients that did not exceed their Shewhart limits are zeroed out while those that exceed the limits are maintained. The reconstructed signal is therefore a series of mostly zeroes and some alarm values. The last step is to monitor the reconstructed series using a Shewhart chart. The limits of this chart are based on the mean of the series and on the standard deviations only of the levels in which there was an alarm. The reconstructed series signals at time t only for points in the reconstructed series that exceed the control limits. This last step of monitoring the reconstructed signal is crucial for extracting the relevant features and for quicker detection of any anomalous behavior in the series [6, 10].

3.3 False Discovery Rate Correction

Since MSSPC includes multiple control charts in parallel, it suffers from an inflated false alarm rate arising from multiple testing. In other words, although the false-alarm rate at each scale is small, the cumulative rate can be very high. In fact, since the scales are orthogonal, the $L+1$ false alarm rate α accumulates to $1-(1-\alpha)^{L+1}$. Classical multiple-comparison procedures such as Bonferroni aim to control the false alarm rate probability in families of comparisons under simultaneous consideration [13]. However, such methods tend to be too conservative in their representation of the false alarm rate. A more powerful method for handling multiple testing is by controlling the *false discovery rate* (FDR) [13]. By definition, the FDR is the expected proportion of false alarms among all of the alarms. We apply an FDR correction to MSSPC (FDR-MSSPC) to account for the $L+1$ tests that take place at every time t. Essentially, this lowers the overall false alarm rate.

3.4 An Application of FDR-MSSPC to the Data

Figure 4 shows the final stage of FDR-MSSPC applied to the GI cc series. Using a no-outbreak period of April 10, 1998 to November 15, 1998, the detection algorithm displays three (3) major outbreak periods, one for each year. The three major outbreaks manifest during the following periods: (1) December 22, 1998 - February 7, 1999; (2) December 9, 1999 - March 13, 2000; (3) November 25, 2000 - February 28,

2001. Within these three major yearly GI cc outbreaks, it is important to note that these ranges incorporate days that, for the most part, are consecutive.

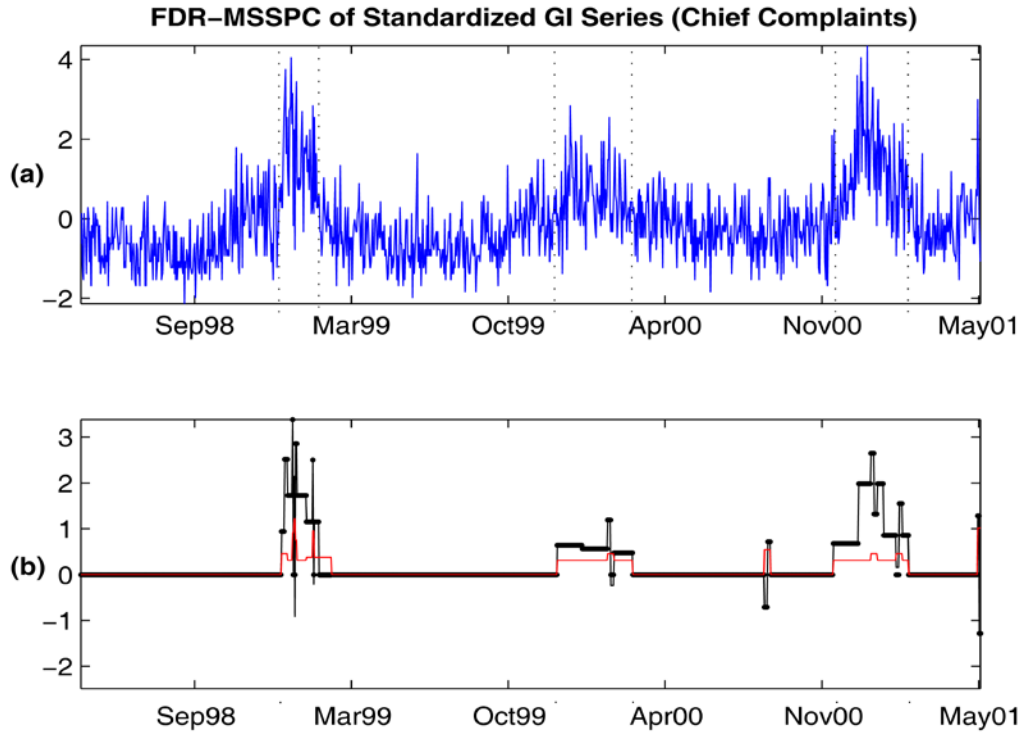

Figure 4: *Reconstruction Stage of FDR-MSSPC applied to Gastrointestinal Chief Complaint Series*. Panel (a) represents the original series. Dotted lines summarize major outbreak detection periods for each year. Panel (b) represents the reconstructed series in the form of dots. Points outside of their own detection limits become alarm values and contribute to the range in (a). Both panels display the same information.

There exist, however, periods in the ranges where some gaps surface but only for small time periods. For example, within the first outbreak period (December 22, 1998 - February 7, 1999), the detection dates produced by the algorithm include January 5th then January 7th then January 9th (as opposed to all dates between January 5th and January 9th). It makes sense, however, to include the range of January 5th - 9th in the detection range since detection dates are only separated by a day. After February 7, 1999, the detection date jumps to December 9, 1999, which clearly cannot be considered as part of the first outbreak period. The largest detection gap within either detection range occurs in the third outbreak and spans 5 days (February 14 - 19, 2000).

4. MULTISCALE PRINCIPAL COMPONENTS ANALYSIS

In many cases, there are multiple time series to be monitored. In practice, there is a tendency to monitor the series separately using univariate methods. Such methods, however, cannot capture the interrelations between the different series and abnormalities which might occur in these relationships. Further, abnormalities in multiple series can go undetected if each series is observed separately. Thus, the information contained in the multivariate nature of the data can be crucial for rapid detection and low false alarm rates. Hence, we discuss a wavelet-based method for multivariate monitoring. This detection algorithm, also developed by Bakshi, is an extension of MSSPC called Multiscale Principal Components Analysis or MSPCA [10].

This approach toward multivariate monitoring, which combines the idea of PCA and wavelets, is based on reducing the dimension of the data and then using univariate charts to monitor the reduced series and the residuals [14]. The MSPCA algorithm consists of decomposing each series using wavelet decomposition; PCA is then applied separately to coefficients from all series at each scale in order to reduce the dimensionality. As in MSSPC, with every incoming observation, the process is repeated in a roll-forward manner. Figure 5 provides a diagrammatic representation of MSPCA.

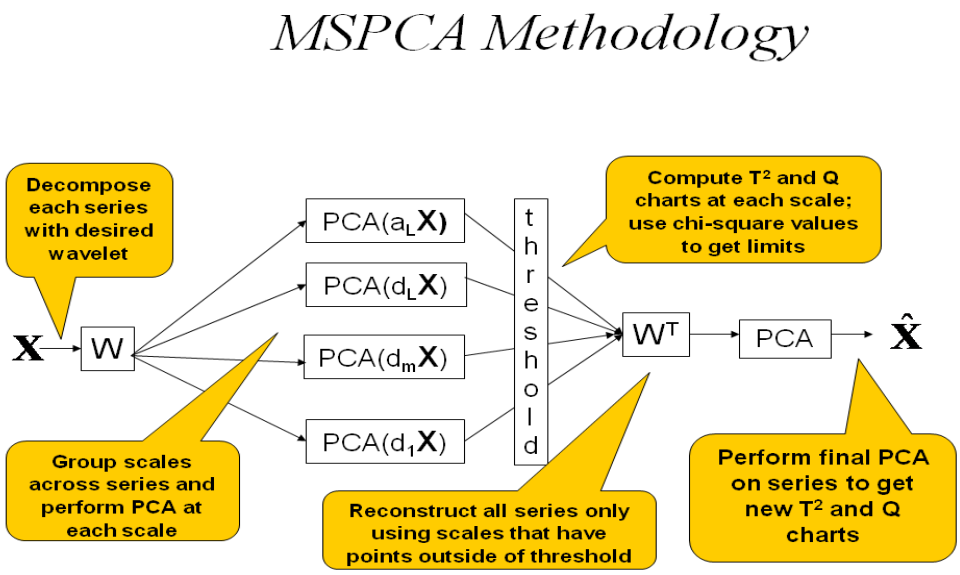

Figure 5: *MSPCA Methodology*

4.1 The Decomposition

The input into MSPCA is a vector of *s* series. At first, each of the series is decomposed using wavelets. Next, principal components analysis is performed on each vector of the detail scales and the finest approximation for all series. The goal is to represent the *s* sets of coefficients at each scale in lower dimension. Assuming that the group of series is not independent, it should be possible to capture their information in a dimension lower than *s*. PCA transforms the *s*-dimensional data into *s* principal components, which are linear transformations of the original data that are not correlated. Usually a small number of principal components (PCs) captures the majority of information concerning variabililty in the original data. The next step is to separate the PCs into *primary* and *residual* PCs. The first *p* PCs contain *primary* information, while the remaining *s–p* PCs house the *residual* information. There are several criteria for determining how many PCs are primary. One is according to the percentage of information they jointly contain of the total information. Other thresholding techniques for selecting the appropriate number of components are the scree test plots and parallel analysis [10].

4.2 T^2 and Q plots

The cluster of primary information from the principal components retained is now monitored by a T^2 chart, whereas the cluster of residual information is monitored by a Q chart. The points on these plots are the sum of the standardized scores. These two sets of plots require the component scores in addition to the corresponding eigenvalues at each level and are summarized in equations (2) and (3):

$$T_{t,j}^2 = \sum_{j=1}^{p} \frac{c_{t,j}^2}{\lambda_j}; \tag{2}$$

$$Q_{t,j} = \sum_{j=p+1}^{s} \frac{c_{t,j}^2}{\lambda_j}, \tag{3}$$

where $T_{t,j}^2$ and $Q_{t,j}$ are the sum of squares of the selected scores ($c_{t,j}$) scaled by the respective eigenvalue (λ_j) computed from the data at the tth time point and jth level [10]; p is the last principal component retained in the *Hotelling-*T^2 (T^2) chart, and s is the total number of series and total number of principal components.

4.3 The Detection Limits

Since the monitored statistic is a sum of squared values and thus always non-negative, we have that $LCL = 0$. The UCLs for the T^2 charts at each level are given by $\chi^2_{1-\alpha,p}$ values, while the UCLs for the Q charts at each level are given by $\chi^2_{1-\alpha,s-p}$ values. A statistically significant change at a certain scale is indicated if either the T^2 or the Q chart at that scale triggers an alarm. Scales for which values do not exceed the χ^2 limits are discarded because this suggests that the information at those T^2 and Q levels contain no relevant information concerning abnormal operation of the monitored process.

4.4 Reconstruction and FDR

Reconstruction of the series is computed only from the details and approximation coefficients whose values exceed the UCL limits. Though there are still s series in this reconstruction, they do not contain the extraneous information found in the decomposition levels stated earlier. Finally, this new reconstructed series vector is subject again to PCA, and its components are again separated into *primary* and *residual* clusters. A final set of T^2 and Q charts is used to monitor the reconstructed series with the same χ^2 thresholds stated previously. This last step of selecting the scales that indicate significant events, reconstructing the signal, and computing the scores and residuals improves the speed of detecting abnormal operation and reduces false alarms [10].

As in MSSPC, it is also needful to address the issue of multiple testing and related false alarm rates in MSPCA. At each time t, there are $2(L+1)$ simultaneous charts, thereby creating a multiple testing situation. The overall false alarm rate is therefore inflated. We again suggest improving this by using an FDR correction to MSPCA (FDR-MSPCA). The FDR correction lowers the overall false alarm rate that would otherwise be $1-(1-\alpha)^{2(L+1)}$.

4.5 An Application of FDR-MSPCA to the Data

For this application, we consider four standardized pediatric series for multivariate monitoring. The first series is the gastrointestinal (GI) cc series, monitored in Section 2. The other three are as follows: a hospital admissions series for GI illness for the same group, and two series (1 for cc and the other for hospital admissions) which track respiratory illnesses in the same youth group. All series are

based on the same focus group (children under age 5) living in the four-county Utah territory known as the Wasatch Front area [11]. It becomes of interest to examine the FDR-MSPCA detection algorithm using these 4 series.

We again use the no-outbreak period as described before: from April 10, 1998 - November 15, 1998. In the algorithm, we retain principal components of 75% for details and approximation coefficients. In other words, 75% of the variance from the principal components are found in the T^2 chart, while the remaining 25% are plotted as residuals in the Q chart. As one might expect, there are 3 primary detection ranges existing in the final T^2 and Q charts, all of which exceed the χ^2 threshold. The

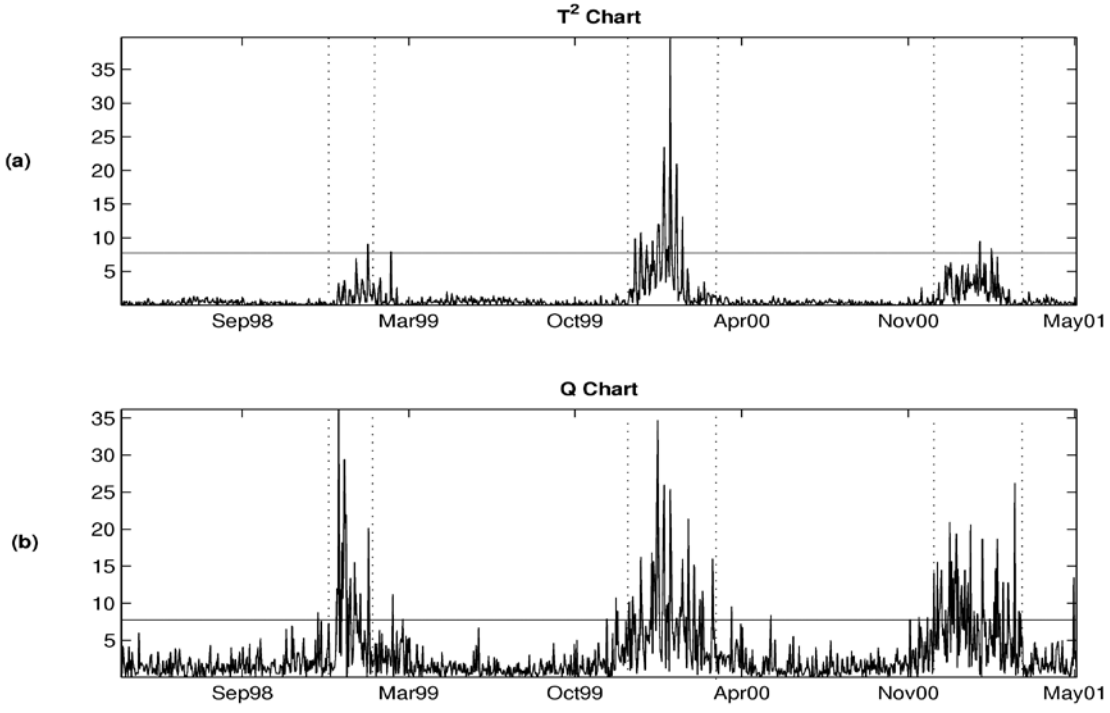

Figure 6*: T^2 and Q Graphs for FDR-MSPCA applied to 4 Pediatric Series.* Panel (a) represents the T^2 graph. Panel (b) represents the Q graph. Dotted lines summarize major outbreak detection periods for each year. Both panels display the maximum detection range per detection period.

three ranges are: December 25, 1998 - February 1, 1999; November 25, 1999 - March 9, 2000; and, November 23, 2000 - March 18, 2001. Figure 6 illustrates these detection dates in each panel. The graph highlights the maximum detection dates for each period in each panel. For example, the second detection range (November 25, 1999 - March 9, 2000) occurs primarily in the Q panel. A subset of it, however, appears in the T^2 panel for the same time points. The dotted lines that appear for this second range

for both panels, then, manifests as the maximum range which exists in the T^2 and Q panels for those particular time points.

Detection ranges in Figure 6 are logical groupings of detection points. As in FDR-MSSPC, each range may not necessarily contain all consecutive time points, but the detection clusters are sensible. For example, the detection range between the second detection range and the third detection range is over the course of months (March, 2000 to November, 2000). This wide-range span makes it obvious that this gap separates detection ranges. Although the within-range detection varies per detection cluster, a natural grouping of detection surfaces. For example, even though the third detection range includes dates which are listed as February 8th and 9th and then as the 16th, the overall detection date spread for this range suggests that it is not abnormal to include the 16th as a bona fide detection point. The output indicates that the natural stopping point for this range is March 18, 2001, given that the next detection point after it occurs two months later.

5. RESULTS

Our results are first based on applying the EWMA technique to the data described in Section 3.5. We use a smoothing constant of 0.2 ($\lambda = 0.2$) since this is what the authors use in the Ivanov paper [11]. We seek to determine earliest detection dates using this monitoring method. Detailed Matlab results for this method are viewable in Figure 7. Using this method and a 3-sigma limit threshold, we notice that the earliest anomaly detection date for each year occurs around Christmas.

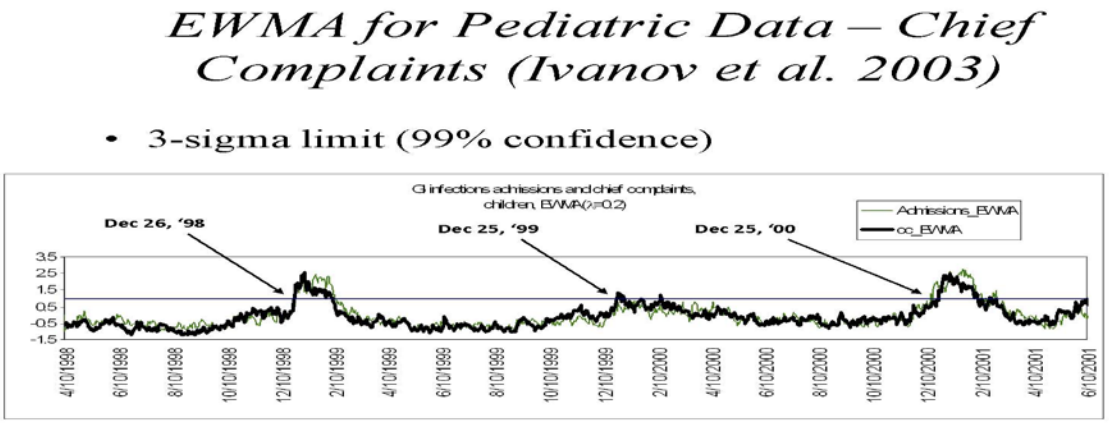

Figure 7: *Monitoring Results Using EWMA.* Using $\lambda = 0.2$, the results show earliest detection dates for each year (1998-2000) for the pediatric data (chief complaints).

A comparison of these results with those of the two wavelet-based methods (FDR-MSSPC and FDR-MCPCA) follows. Table 1 provides a summary snapshot of earliest detection dates or points which fall outside of the control limits by method.

Table 1: Earliest Detection Dates by Method

Year	*EWMA*	*FDR-MSSPC*	*FDR-MSPCA*
1998	Dec 26	Dec 22	Dec 25
1999	Dec 25	Dec 9	Nov 25
2000	Dec 25	Nov 25	Nov 23

For both wavelet-based methods, we have guarded against multiple false alarm rates; for MSPCA, detection dates in the table reflect interplay between four multivariate syndromic data streams. With this dataset, both wavelet-based methods outperform the more traditional EWMA method. The two wavelet-based methods provide earlier detection dates for each of the three years. For 1999 and 2000, MSPCA signals an alarm an entire month before EWMA does. These methods become key, as the aim is to provide the earliest anomaly detection, which translates into early action in order to save lives.

6. FINAL DISCUSSION AND CURRENT CHALLENGES

In this treatment, we examine wavelet-based methods for monitoring frequent data related to adult and pediatric GI ailments. MSSPC and MSPCA require less pre-processing and assumptions than other traditional statistical monitoring methods; however, there are a few parameters that must be specified by the modeler. The first is the choice of wavelet. Here, we have used the Haar because this wavelet is useful for detecting sudden, abrupt changes in time series. In biosurveillance, this is significant since the goal becomes to monitor sudden spikes in the data, which

translates into peaks in sales of syndromic data. Since the Haar is a wavelet whose basic function is to average and subtract consecutive points in the series, it detects quickly any two consecutive points that have a large range. For gradual, slower changes in the series, however, a different wavelet would have to be employed in order to make monitoring more effective.

Second, the modeler must specify the number of decomposition levels to be used in the analysis. In our analysis, we use five levels ($L = 5$). In general, the number of levels depends on the wavelet used and the amount of notable information captured at each level. The analyst is required to investigate the results and determine L for each separate trial.

We conclude that, when compared to EWMA, wavelet-based detection algorithms provide earlier detection of outbreaks for gastrointestinal illness in children less than five years old in the Utah counties examined. In addition, these techniques are computationally efficient and user-friendly: the wavelet toolbox in Matlab makes the run time of the algorithms a rather quick process, and the output are a nice set of graphs and a list of detection dates. The ease of interpretation of the results and speed of the software make these methods particularly useful in public health circles, where time is of the essence for early intervention.

One main obstacle in the advancement of this field is that the relevant literature that exists spreads so vastly across varying fields [15]. Although more conferences of late have begun to focus on anomaly detection in biosurveillance and in disease outbreak, researchers must continually strive to put forth a concerted effort to document pinpointed statistical methods used in practice for this relatively new field.

The major challenge with analyzing wavelet-based methods in health care is data acquisition. The pressing question is, "How can statisticians gain more steady access to much of the data which is classified as syndromic?" [15]. Unless statisticians are somehow associated with researchers in "syndromic laboratories" or have some other inroads to inquiring this specific type of data, these cutting-edge methodologies will be difficult to explore. Although it is understandable that confidentiality of such data is paramount, there must be a better meeting of the minds between those in industry and academia if new methodologies (wavelet-based or otherwise) are to be adequately tested.

References

1. Wagner M, Robinson JM, Tsui F, Espino J, Hogan W. Design of a National Retail Data Monitor for Public Health Surveillance. *Journal of the American Medical Informatics Association* 2003; **10**:409-418.

2. Fienberg SE, Shmueli G. Statistical Issues and Challenges Associated With Rapid Detection of Bioterrorist Attacks. *Statistics in Medicine* 2005; **24**:513-529.

3. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales. *Proceedings of the National Academy of Sciences* 2002; **99**:5237-5240.

4. Zhang J, Tsui F, Wagner M, Hogan W. Detection of Outbreaks from Time Series Data Using Wavelet Transforms. *Proceedings of the AMIA Annual Symposium* 2003; pp. 748-752.

5. Shmueli G. Wavelet-based monitoring for modern biosurveillance. Technical report, RHS-06-002, University of Maryland, Robert H. Smith School of Business.

6. Aradyhe HB, Bakshi BR, Strauss RA, Davis JF. Multiscale Statistical Process Control Using Wavelets – Theoretical Analysis and Properties. *AIChE Journal* 2003; **49(4)**:939-958.

7. Graps A. An Introduction to Wavelets. *IEEE Computational Science and Engineering* 1995; **2(2)**:50-61.

8. Percival DB, Walden AT. *Wavelet Methods for Time Series Analysis*. Cambridge University Press: UK, 2000.

9. Donoho DL, Johnstone IM, Kerkyacharian G, Picard D. Wavelet Shrinkage: Asymptopia. *Journal of the Royal Statistical Society, Ser. B* 1995; **57(2)**:301-369.

10. Bakshi BR. Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE Journal* 1998; **44(7)**:1596-1610.

11. Ivanov O, Gesteland PH, Hogan W, Mundorff M, Wagner M. Detection of Pediatric Respiratory and Gastrointestinal Outbreak from Free-Text Chief Complaints. *Proceedings of the AMIA Annual Symposium* 2003; pp.318-322.

12. Hubbard BB. *The World According to Wavelets*. A.K. Peters, Ltd.: Wellesley, MA, 1996.

13. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 1995; **57**:289-300.

14. Shmueli G, Fienberg S. Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance, chapter in *Statistical Methods in Counterterrorism*, Springer: New York, 2006.

15. Shmueli G, Burkom HS. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics (Special Issue on Anomaly Detection)*, in press.