Framework for using grocery data for early detection of bio-terrorism attacks[a]

Anna Goldenberg
October 2001
CMU-CALD-01-101

Center for Automated Learning and Discovery
Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh, PA 15213
anya@cmu.edu

Abstract

Early detection of epidemics and bio-terrorism attacks is of great concern to public health. There are many sources of data that might be used for early detection. The main goal of this project is to investigate the possibility of detecting epidemics and bio-terrorism early by analyzing trends in consumer grocery purchases. This type of data has two main advantages: first, we expect grocery data to have an earlier signal of an outbreak, since people tend to seek self treatment of symptoms before they reach a doctor or a hospital. Second, grocery data are much richer and are available on a more refined scale than epidemiological data, that were previously the main source of information used for detection purposes. This paper introduces a framework that combines techniques from signal processing, forecasting and quality control to increase the sensitivity of the system and the rate of detection. Over-the-counter medication purchases were extracted from grocery data and used for experiments. This paper shows that grocery data can be used for the timely detection of epidemics and bio-terrorism attacks. The automated framework developed here is a first step towards a new generation of systems based on non-specific syndrome data, i.e. data collected for purposes other than medical analysis. There is a hope that this work will have significant impact on the evolution of early detection and consequently on computer-based surveillance (CBS) systems.

[a]Matlab code for Automated Framework is available at http://cs.cmu.edu/~anya/BTproject/code/

1 Introduction

Many threats to the public health are biological in nature. Some of them occur and mutate naturally, causing various epidemics, and others are engineered and introduced into the environment intentionally in order to harm the population. The effect of introducing a bio-agent can be enormous bringing deaths to many people and thus requires identification and proper response as soon as possible.

1.1 Motivation

Bio-terrorism and its ability for mass destruction have been of increasing concern to the American goverment and internationally. The recent events of anthrax exposure have underscored the importance of bio-terrorism awareness as never before. A threat that only a few weeks ago seemed hypothetical has become a reality. As of October 14th seven people in three different states are known to have been exposed to anthrax, as reported by Philip Shenon from New York Times on October 15th, 2001. Only two of them however were ill, including the death of the 65 year old man from Florida. Though the source of anthrax has not been established yet, "It clearly is an act of terrorism to send anthrax through the mail," said Health and Human Services Secretary Tommy G. Thompson on Fox News. Mr Thompson had also stated that the White House would ask Congress for an additional $1.5 billion for the purchase of the antibiotics and for other programs to combat bioterrorism, as reported by New York Times.

A known case of bio-terrorism has been recorded in 1984 in Dalles, Oregon. The Rajneeshee religious cult planned to infect residents with Salmonella on the election day to influence the results of county elections (Torok et al 1997). To practice for the attack, they contaminated salad bars at ten restaurants with Salmonella Typhimurium on several occasions before the election. A community-wide outbreak of salmonellosis resulted; at least 751 cases were documented in a county that typically reports fewer than five cases per year. Although bioterrorism was considered a possibility when the outbreak was being investigated by public health officials, it was considered unlikely. The source of the outbreak became known only when the FBI investigated the cult for other criminal violations. A vial of Salmonella Typhimurium identical to the outbreak strain was found in a clinical laboratory on the cult's compound, and members of the cult subsequently admitted to contaminating the salad bars and putting Salmonella into the city's water supply tank. This incident, among other recent events, underscores the importance of improving preparedness at all levels.

A group of scientists from the University of Pittsburgh has studied the nation's current capacity for the early detection of public health threats including bio-terrorism (Wagner et al, 2001). They have identified 66 systems that were either a part of the national servaillance network or stand-alone systems. These systems work with various types of data, such as symptom information, disease reports, culture results obtained from hospitals, clinics, labs and emergency rooms. There are several characterists that most systems have in common. Since most of the information comes from "medical" sources, there are bound to be delays in obtaining and reporting the data.

Most people do not call emergency services immediately but start by taking pain killers, looking for symptoms on the web, calling friends, etc. By the time the medical community has received a substantial amount of calls to get concerned - it may be too late.

It is desirable to detect an epidemic or an attack as soon as possible, but what are the sources that first notice the increase of the public health threat? Where is the information that could hint the first signals of sickness in a large group of people? Traditional "medical" data is a valuable source, but contains delays, so the solution is to use data collected for purposes other than

epidemic/attack detection, in other words, non symptom-specific data.

Several potential sources were looked at. School absenteism was one of them. Unfortunately, schools are out during holidays and breaks, hence it cannot be a reliable source of data all year round. Internet information was of great interest, however only several specific sites are collecting personal information such as users locations. Other general purpose sites such as search engines Google and Lycos do not have user specific information, hence the web data cannot be used to detect epidemic/attack at the regional level.

When people recognize symptoms of a disease that they have already experienced before, they might attempt self-treatment first. For example, if a person has fever, he/she might take Tylenol or another fever reducing medication. If the symptom is cough, Nyquil might be their choice. Hence, over-the-counter medications and other groceries, such as orange juice or Kleenex may be a valuable timely source of information. This data is very rich, and although it does not measure illnesses directly, it can infer specific symptoms that are being experienced by the purchasers at a relatively early stage of the outbreak. Although this data is typically noisy, the potential of discovering the first signals of an outbreak (or a bio-terrorism attack) is promising. Goldman (2000) showed that the potential of such data for the detection of seasonal flu, that in turn might be an indication of a bio-terrorism attack, can be enormous.

1.2 Paper Overview

In this paper we show that it is feasible to use non symptom-specific data, such as over-the-counter medication sales, for early detection of epidemics and bio-terrorism attacks. The advantages and complications of utilizing non-symptom specific data as well as the intricacies of grocery data are described in Section 2. An extensive and flexible framework proposed in this paper can be implemented at the supermarkets or pharmacies.

Very little is known about the ways epidemics manifest themselves in grocery data. Even less is known about bio-terrorism attacks. Hence, the system could not be constructed using an ordinary supervised learning approach. Our framework gains its sophistication from using a combination of methods from various disciplines, such as time series analysis, forecasting and statistical quality control. The system described is flexible so that higher accuracy predictive methods could be readily substituted into the appropriate stage of the system. Each of the stages of the framework is described in detail in the Framework Description section.

Since the information about epidemics or bio-attacks in grocery data is not available, the ordinary measures of sensitivity, timeliness, and specificity can not be used to assess the performance of this detection system. Evaluation methodology is described in detail in the Evaluation Section 4. Discussion of the strength and weaknesses of the proposed early detection system (Section 7 and ideas for future research (Section 9) conclude this paper.

2 Data Description

The reasoning behind using grocery data for detecting epidemics and bio-terrorism attacks is simple. Most, people when they have a mild headache, or sore throat start by taking medications. If it is a mild muscle pain or light fever, they might take Tylenol. General purpose medications, like Tylenol, may be available in supply at home. However if the symptoms are more specific, like conjested throat or stuffed nose, they might want to have a throat spray or nasal inhaler that come in small containers. This would require them to go to the store and buy the appropriate

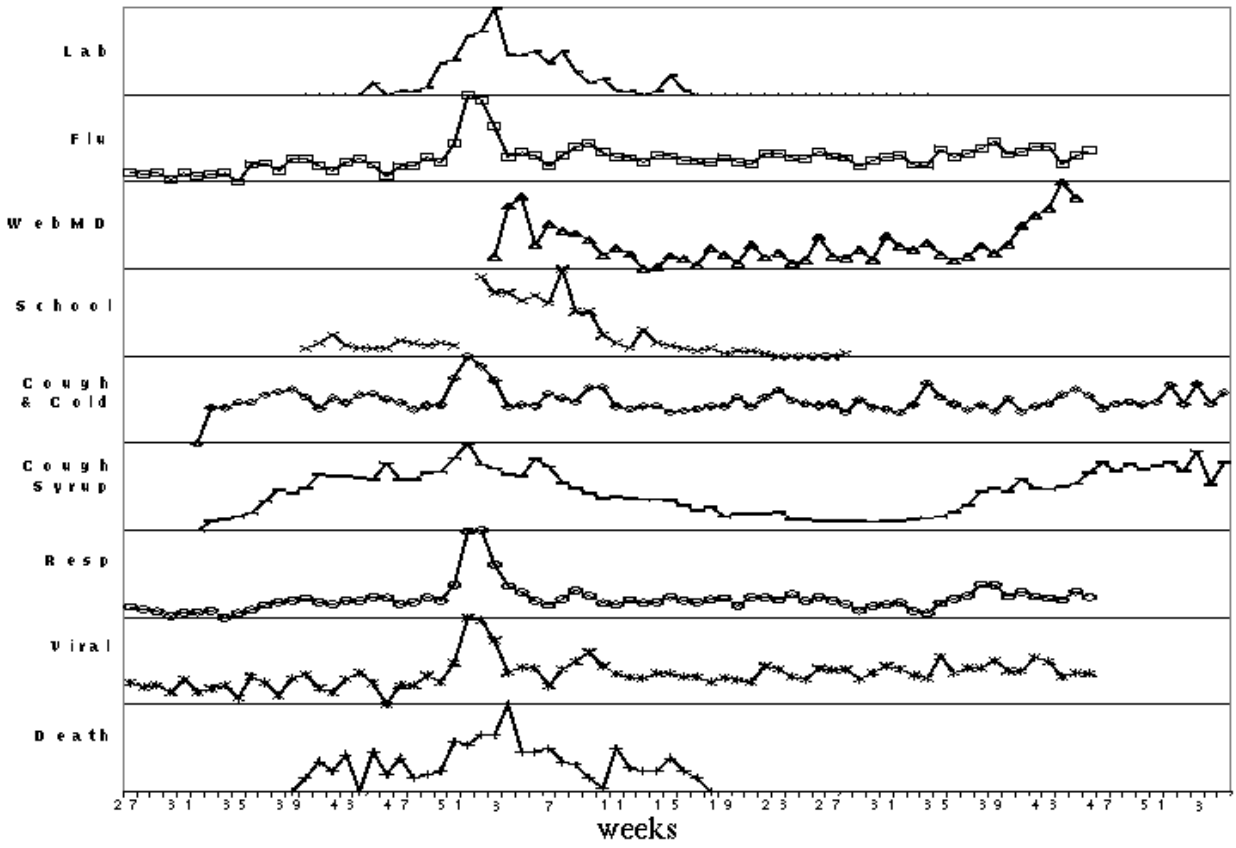

Figure 1: Weekly counts for several types of routinely collected data for different time periods around the December 1999 Influenza outbreak in Pittsburgh. Each data type is plotted on a normalized 0-1 scale. Legend: Lab, influenza cultures from the UPMC Health System; Categories of emergency department ICD-9-coded chief complaints; WebMD, counts of queries to a national web health site using words such as fever and cough; School,school nurse influenza reporting; Cough&Cold and Cough Syrup, weekly counts of over-the-counter medication sales; Resp. and Viral, collection of the lab test results; Death from Obituary Records.

medications. Only if symptoms were sudden and severe or the traditional self-treatment was not working, people would report to their doctor or even call emergency services.

It is not just rational thinking that supports usefulness of grocery data usage for early detection. New York's Health Department has investigated the usage of pharmaceutical sales for the detection of diarrheal diseases (Mikol, et al 1999). It was established that medication sales information were a valuable source for early detection.

Correlation analysis of several datasets, specifically comparing grocery data to medical data that is currently used in surveillance systems, was conducted as part of collaborative effort of scientists from Carnegie Mellon and Pittsburgh Universities. Several datasets used for comparison are depicted on Figure 1, which is reproduced with permission from Wagner et al (2001) with addition of the grocery datasets. A number of the datasets represented on the figure, such as flu, respiratory and viral tests, and the lab culture reports, are medical datasets currently used for surveillance of epidemics and attacks. Death records and school abcenteism are non-medical datasets that are used in serveillance systems of several states. WebMD may prove to be a valuable

dataset in the future however the data was not available in sufficient quantities for analysis. Two grocery datasets are represented on the graph as well: The Cough&Cold dataset is a group of medications that consists of several subgroups, such as cough lozenges, denoted by cough/syrup on the graph. From Figure 1 it is clear that all series have an upward spike around week fifty one. Since the resolution of the x-axis represents is weeks, the timeliness of the grocery data is not evident. However, preliminary research suggests that alarming signals may appear in grocery data three to four days prior to appearing in the traditional medical data.

There are certain advantages to using grocery data, of which timeliness was already discussed. Another advantage is the level of details available. Each grocery purchase is recorded in every store of the grocery chain. Extensive information about the exact time and place of each purchase is available. With the introduction of advantage cards, information about buyers households may also become accessible. Knowledge about the details of the purchases may become very important, for example in identifying locations if the attack was targeted for a particular region.

There are however, drawbacks, to using grocery data. The manifestations of bio-terrorism attacks in grocery data were not studied widely. The identification of abnormalities is further complicated by outliers due to sales patterns. Also, distinguishing between epidemics and bio-terrorism attacks might present difficulties. This calls for further study by epidemiologists.

Data used in the project are daily retail over-the-counter medication sales. The information is obtained by summation of sales for individual products over the period of a single day from a number of stores of a grocery chain in the Allegheny County. This means that the information of sales of combinations of products such as basket information is not available. There were 541 daily datapoints available starting August 8, 1999 and ending January 31, 2001 for several different product categories such as Cough Syrup/Liquid Decongestant; Tabs and Caps (including Advil Cold/Sinus, Tylenol Flu Non-Drowsy, etc); Throat Lozenges/Cough Drop and Nasal Spray/Drops Inhalers. Some of the subgroups are shown in Figure 2.

The datasets selected for the project and represented on the graph are over-the-counter medications used for the treatment of inhalational diseases, since the project stemmed from an effort to identify anthrax that is known to have flu-like symptoms. It can be seen from Figure 2 that all datasets have at least one feature in common: sales seem to be higher in winter than in summer. It is expected since people are more susceptible to catching the flu during winter.

Most of our experiments were carried out using the "cough" dataset. The counts for cough sales per day can be found on Figure 3.

This series exhibits a summer/winter trend. It is also possible to observe big peaks around winter holidays such as a huge spike between Christmas and New Year. There is also a drop down almost to zero around April 23rd, which indicates that most stores were closed on Easter. Other expected trends include a weekly 7-day periodicity, representing the fact that the sales during the week are generally lower than on weekends and that the highest sales day is usually on Saturday.

To study periodicities present in the data, we introduce a new graphical tool that removes seasonality components iteratively leading to a *periodicity plot*. The idea is to remove one seasonal component from the data at a time. The first component is every other day ($i = 2$), then every third day ($i = 3$), etc. After each component is removed, the variability of the data is measured and plotted as a point on the periodicity plot. We expect that the removal of a component that exists in the data will reduce the variability considerably, while the removal of a periodic component which does not exist will only reduce the variability of the data slightly. The actual removal of a certain component is done by averaging all data points that are of that periodicity (i.e. every ith day), and subtracting this average from each of the above points $(i, 2i, 3i, ...)$.

The periodicity plot then visually implies which substantial periodic components exist in the

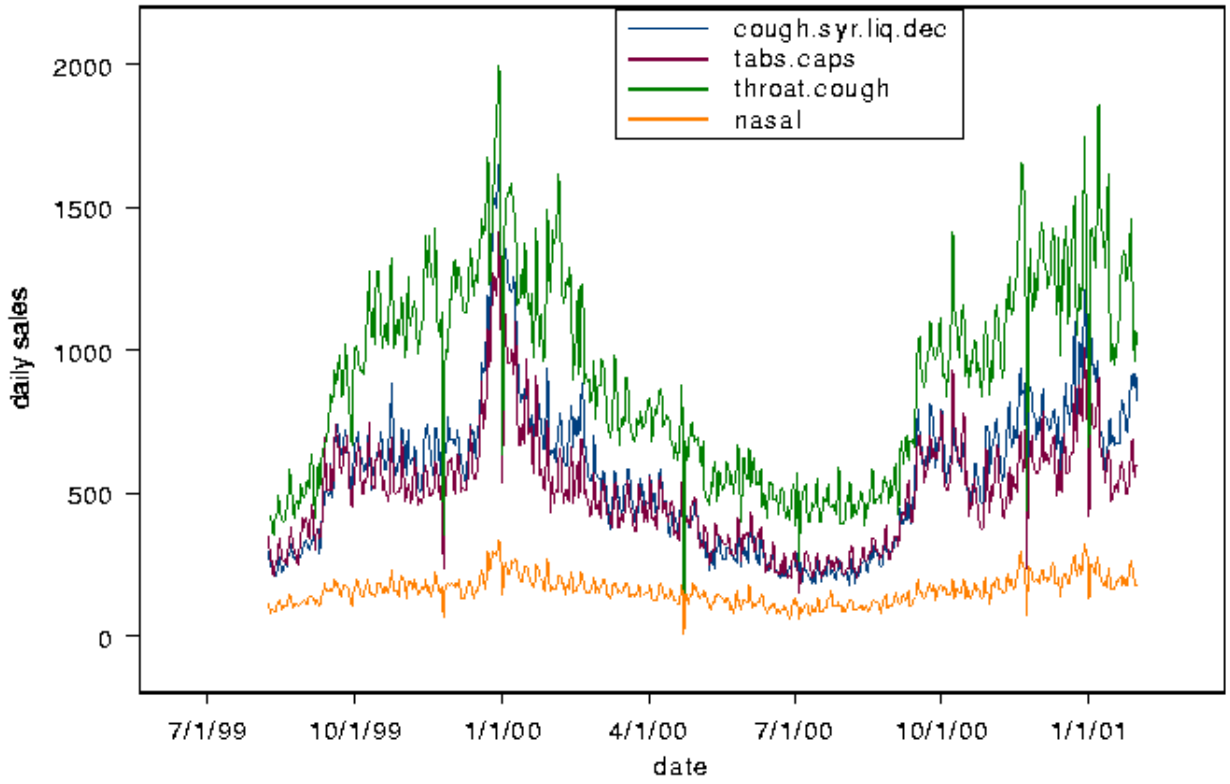

Figure 2: Sample of grocery datasets

data. These components can then be interpreted, and if sensible, removed. Figure 4 gives the periodicity plot for the scaled cough medication sales. It is clear that there is a strong 7-day component which manifests itself in sharp dips in the plot on periodicities with multiples of 7.

We can see that the removal algorithm is effective by inspecting Figure 5 where the same dataset is shown without 7-day periodicity, removed by the method described above.

Even though the periodicity removal may be successful, it is not clear that the single removal of 7 day periodicity is helpful in increasing predictability of the dataset. Therefore, a variation of data pre-processing and prediction techniques were used to test the framework.

3 Framework Description

Very little is known about the ways epidemics manifest themselves in non-symptom specific data. Even less is known about bio-terrorism attacks. Hence, the system could not be constructed using supervised learning (Mitchell, 1997). The goal of the project was to propose an early detection framework that could be easily implemented in any non symptom-specific environment where data becomes available on a frequent, for example day-to-day, basis.

The developed system could be roughly described as follows:

1. Train the model based on non-epidemic regular data previously observed by the system.
2. Predict the number of purchases of a certain product for the next day.
3. Select a threshold that would serve as an upper bound for regular sales behavior.

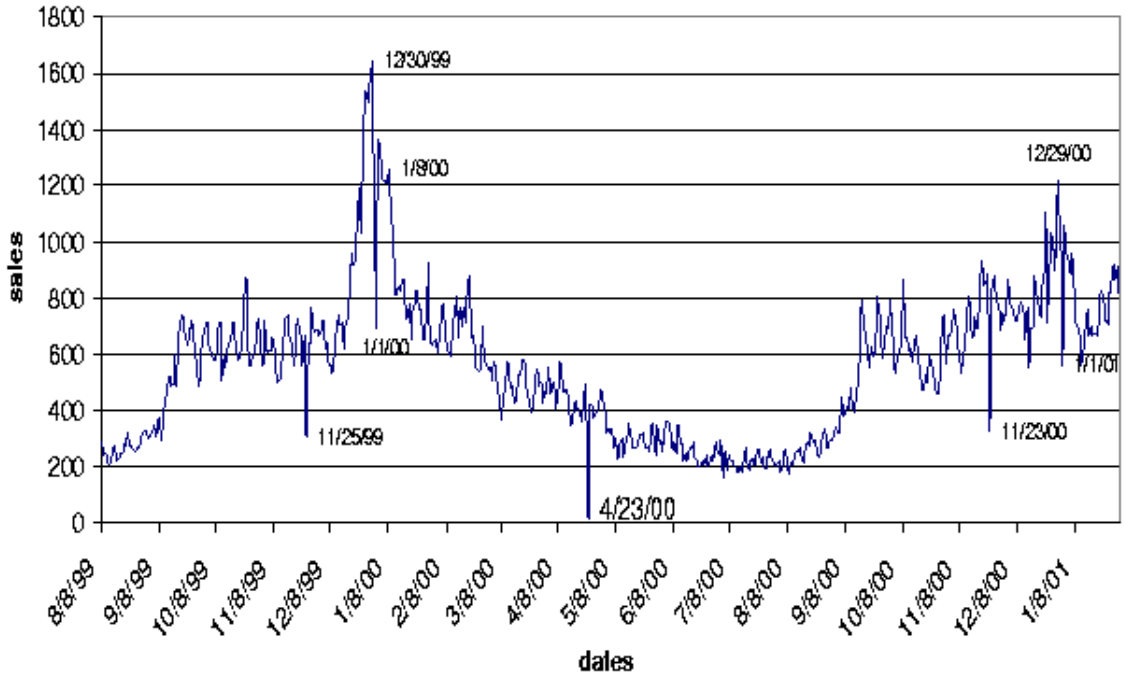

Figure 3: Cough Syrup/Liquid Decongestant counts

4. If the new daily sales volume exceeds the threshold, then raise an outlier flag.

5. Proceed repeating steps 1 through 4.

The outline above is meant to give a very general overview of the system. Due to the complexity of the data, the task of reliably predicting next-day sales is challenging calling for a complicated system within itself.

It was essential that the framework be well-structured and easily traceable, hence the preference was given to simpler models as opposed to complex "black-box" algorithms, such as neural networks. The framework is composed from a variety of methods taken from different scientific fields, such as signal processing, time series analysis, forecasting tools, etc. It is important to note that the techniques used in the framework are not novel. However, they are applied to a novel problem and their combination is also unusual.

The rest of this section is dedicated to describing each of the levels of the framework in detail according to the following schema:

1. Normalizing or scaling the data

2. De-noising the Data

3. Predicting the next day sales volume

4. Selecting a threshold based on historical data

5. Comparing the real sales volume to the threshold

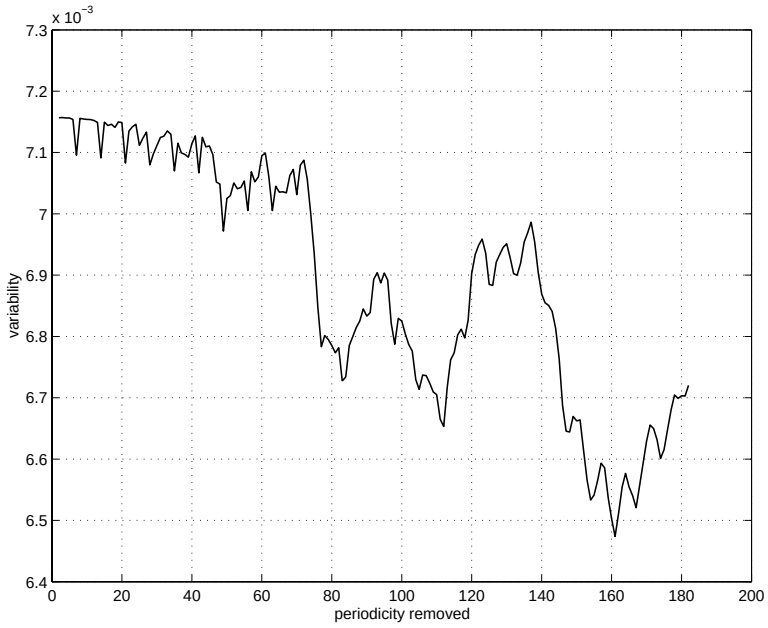

Figure 4: Periodicity plot for the scaled cough medication sales

6. Raising a flag if the sales volume exceeds the threshold

3.1 Data Preprocessing

The framework is designed to work with non-symptom specific data, such as medication sales volumes at a retail store. The data itself is highly periodic. It includes information about day-to-day variations, such as sales on Saturdays being higher than sales on Wednesdays, information about high increases in sale volume during holidays, etc. Daily and weekly variations may be due to the sales patterns rather than the fluctuations of disease in the population, hence they should be eliminated as much as possible for the purposes of this project. One way to reduce the variability due to seasonal effects and to suppress seasonality in the data is to scale the sales within a category by the total daily sales of all products. The store-wide sales information was not available and we used Health group sales, which is an approximation to the total sales. The scaling of the data is roughly equivalent to averaging the sales effects in the data.

Another option which we did not pursue was to use the sales of a relatively stable item as the normalizing constant. However, this type of scaling has two disadvantages: if the normalizing sales series is stable because of its insensitivity to sales patterns, using it for scaling is equivalent to dividing by a constant. On the other hand, by taking a ratio of the sales of two products, a signal in either series can be masked by the other. For those reasons we chose to scale by total sales as described by:

$$\text{normalized data} = \frac{\text{daily counts}_{cough}}{\sum_{\forall categories} \text{daily counts}} \tag{1}$$

Due to the nature of the data, there were several data points in the grocery dataset that were nearly zero. They indicate store closing days, such as Easter. We replaced these zero values with

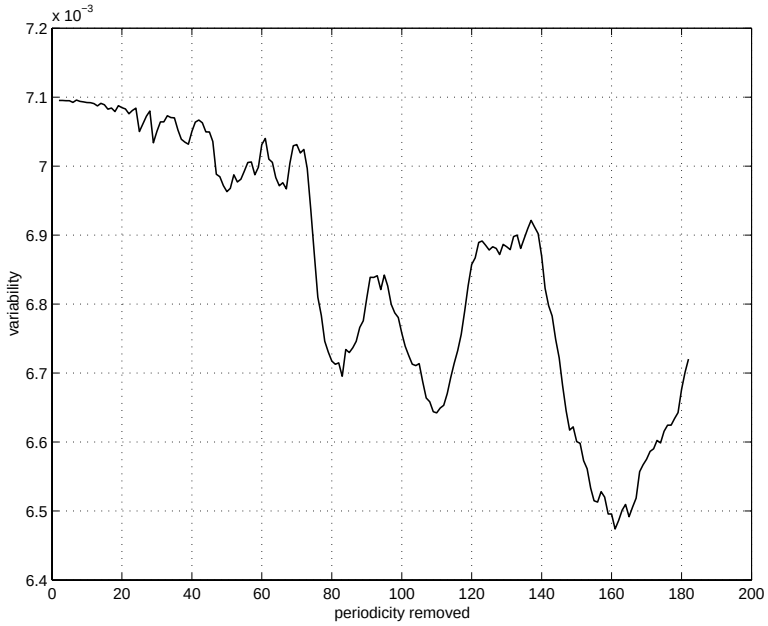

Figure 5: Periodicity plot for the scaled cough medication sales without 7-day periodicity

values obtained by interpolation of four points prior and four points after. The information on the days with interpolated counts should not be trusted.

3.2 De-noising Data

The task of creating an early detection system that is designed to improve public awareness and public health requires high reliability and high accuracy. Noisy and chaotic data is very hard to predict, hence the forecasts made directly from raw or even scaled data are not very reliable. One of the preprocessing steps is therefore targeted to de-noise the data, i.e. to eliminate effects of irrelevant noise in the data thereby making the data easier to predict.

There are many ways to de-noise data. A lot of techniques, such as Fourier transforms and wavelets, are borrowed from the signal processing field for the purpose of noise reduction. Each of the methods has its own merits and backdraws. It is important that effects that are rare and irrelevant, i.e. noise in the data that is weak and infrequent is eliminated. This translates naturally into frequency domain.

In this work we used the Discrete Cosine Transformation (DCT) to de-noise data (for technical information please refer to Appendix A). Unlike most other techniques that are aimed at modeling localized signals, DCT can be used to create as detailed as possible yet general picture of the data. This is done by computing the cosine transformation of an input vector. To reconstruct the original data vector, the Inverse Discrete Cosine Transform is used (see Appendix A).

3.2.1 Coefficient Elimination Technique

In order to decide which DCT components should be eliminated for the purpose of de-noising the data, we apply a technique that eliminates coefficients that have a magnitude below some threshold. For example, setting the threshold to 0.1 will eliminate all coefficients that are of lower magnitude

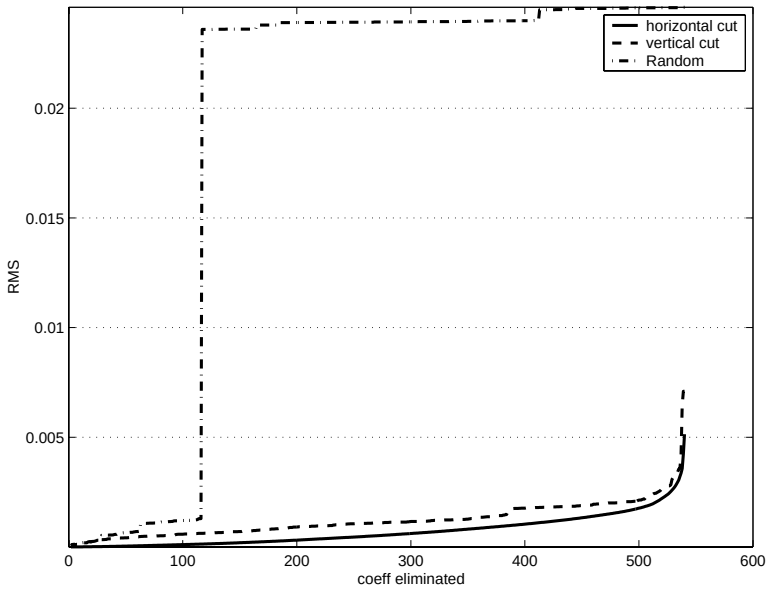

Figure 6: The effect of eliminating DCT coefficients on the RMS(normalized,de-noised) measure for three types of elimination: using a horizontal, vertical, and random filtering value

than 0.1. Naturally, if we set the threshold to 0, the original vector would be reconstructed with 100% accuracy. We call this method "horizontal filtering" to distinguish it from high-pass/low-pass filtering, or "vertical filtering", method that is frequently used in signal processing and eliminates coefficients based on their frequency. Using a horizontal rather than a vertical filter attempts to eliminate weak effects in the data while keeping the de-noised set as close to the data as possible, for the purpose of accurate forecasting.

To compare the effect of removing coefficients on the closeness of the de-noised data (denoted by $D_1, \cdots, D_n$) to the normalized data (denoted by $X_1, \cdots, X_n$), we define a Root Mean Square (RMS) measure for n data points, as:

$$RMS(\text{normalized,de-noised}) = \sqrt{\frac{\sum_{i=1}^{n} (Xi - Di)^2}{n}} \tag{2}$$

The effect of the coefficient removal on the Root Mean Square measure using three different methods of filtering (horizontal, vertical and random filtering) is represented on Figure 6.

The graph is interpreted as follows: Since the data vector can be represented by a vector of DCT coefficients of the same size, if no coefficients are removed, the RMS(normalized,de-noised) is zero. Otherwise, the more coefficients are removed, the greater is the RMS.

From Figure 6 it is evident that the removal of the coefficients using the horizontal cut represented by the solid line has the least RMS of the three methods, for any given number of coefficients used to represent the data vector. The third method, random coefficient elimination, was added as the alternative to show that it does indeed matter which coefficients are removed. It is also possible to notice that the horizontal cut (horizontal filtering) has a less drastic effect on the RMS, meaning that addition or removal of a coefficient results in smaller error than for other methods.

3.2.2 Filtering Value Selection

The threshold for removing DCT coefficients is selected according to predictability and goodness of fit criteria which are determined automatically. The higher the filtering value the more coefficients are removed. The more coefficients removed the simpler and more predictable the de-noised set becomes. If all but the first coefficient are removed, the resulting de-noised set is reduced to a straight line, i.e. easy to predict, but the Root Mean Squared (RMS) of the difference between the normalized data and the de-noised data, is the highest. It is therefore undesirable to select a filtering value that is too high. Also, if the filtering value is too low, the de-noised set would be equivalent to the normalized one and that would defeat the purpose of de-noising. So, to select a de-noised set that approximates the data reasonably, i.e. the filtering value is not too high, and still has high prediction accuracy. The existing data was separated into two parts: a training part of size n_1 and a prediction part of size n_2. The size of the training part is selected to be big enough to obtain a reasonable prediction. We found that a month worth of data is sufficient, i.e. $n_1 = 31$. The prediction part is the remaining data. For example, if 365 points were available in the dataset, then $n_2 = 365 - n_1$. The DCT coefficients are sorted in ascending order, then each coefficient, starting with the highest one is selected to be the filtering value. The de-noised set is obtained by applying IDCT to all coefficients that exceed the filtering value and three RMS error rates are monitored as described in equations $3, \cdots, 5$.

$$RMS(\text{normalized, de-noised}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - D_i)^2}{n}} \tag{3}$$

,where $X_1, \cdots, X_n$ is normalized data and $D_1, \cdots, D_n$ is de-noised data.

Notice that RMS(normalized,de-noised) is measured using all of the available n data points. As before, it represents how close the de-noised set is to the normalized data, i.e. the goodness of approximation. RMS(normalized,de-noised) is expected to decrease as more coefficients are used to obtain the de-noised set. Next, using a one-step ahead linear prediction AR(1) (see Appendix B), the $n_1 + 1$ point is predicted from the de-noised data, and denoted by P_{n_1+1}. We then continue to predict points $n_1 + 2, n_1 + 3, \ldots, n$ using a roll forward one-step ahead prediction, to obtain $P_{n_1+2}, P_{n_1+3}, \ldots, P_{n_2}$. The discrepancy between these predictions and their corresponding normalized values in the prediction set is measured by

$$RMS(normalized, prediction) = \sqrt{\frac{\sum_{i=n_1+1}^{n}(X_i - P_i)^2}{n_2}} \tag{4}$$

$RMS(normalized, prediction)$ measures how well the de-noised set can be predicted with respect to the prediction goal, i.e. real data represented by the de-noised set. It increases as more coefficients are used to obtain de-noised set.

The third measure represents the "predictiveness" of the de-noised data, it is measured by the RMS(de-noised,prediction). The de-noised value for each of the points in the prediction set is computed by applying DCT to all the data until that point. For example, to obtain the $n_1 + 1$ de-noised value, the first n_1 data points are de-noised. The difference between the de-noised values of the prediction set and their corresponding predictions, which were made in the previous step, is assessed by

$$RMS(\text{de-noised,prediction}) = \sqrt{\frac{\sum_{i=n_1+1}^{n}(D_i - P_i)^2}{n_2}} \tag{5}$$

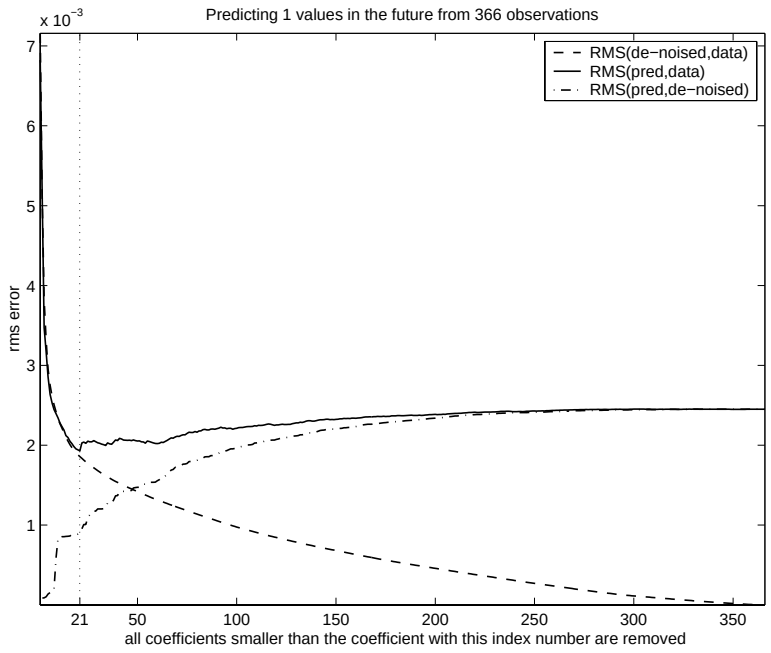

Figure 7: Minimizing the three types of RMS errors for the cough medication sales. The sum of the three is minimized at 21 coefficients.

Unlike the previous two RMS measures, the last is expected to decrease as a function of the number of removed DCT coefficients.

The three RMS error rates are monitored for an increasing number of DCT coefficients and the set of coefficients that minimizes the sum of the three errors is selected as the DCT threshold. This method facilitates the selection of the de-noised set that describes the data best under the constraint of high prediction accuracy. In Figure 7 it can be seen that for the cough medication sales the sum of the three errors is minimized when the number of DCT coefficients (with the largest magnitude) that are retained is 21. The de-noised sales data are then the result of applying IDCT to those 21 coefficients.

3.3 Next Day Sales Forecasting

There exists a wide variety of forecasting models. In general, they are divided into two categories: linear and non-linear predictive models. Linear models represent an observed series as a linear function of the present and the past values of a purely random process (Chatfield, 1989). They are simple to understand and interpret, straightforward to implement, and have been predominant forecasting tools for more than 50 years (Gershenfeld and Weigend, 1993). Autoregressive (AR) and Moving Average (MA) models are included in this category, as well as the more general ARMA, ARIMA (Auto Regressive Integrated Moving Average) and sARIMA (seasonal ARIMA) models. The main disadvantage of linear models is that they are inappropriate for modeling even moderately complicated series, and have encountered various limitations in real applications (Peña et al., 2001). Non-linear models, such as bilinear models and Threshold Autoregressive (TAR) models, are more flexible, and can be used to describe real generating processes that are non-linear (i.e. they allow for a non-linear function to express the relation between the observed series and the past and present values of a random process). The price for this flexibility is the complexity of the model and its

statistical properties (for a detailed discussion, see Chatfield 1989).

From preliminary experiments with grocery data, we learned that linear models such as ARIMA, when applied to normalized data, are not able to capture the complexity of the data's structure and provide poor predictions. In addition, whether applying linear or non-linear models, the underlying assumption for most models is that the series are stationary. A stationary series is one that contains no systematic change in the mean (a trend) or variance, and strictly periodic variations have been removed (for a mathematical definition of stationarity, see Brockwell & Davis, 1996). When the non-stationary features are not of primary concern, transformations can be used in order to achieve stationarity, such as differencing. This is another reason for not fitting a linear model to the scaled or de-noised data directly: since the pre-processing that is needed in order to achieve stationarity is data-specific, it would be hard to incorporate it into an automated system.

We therefore introduce an additional layer that is suitable for non-stationary series, and that improves the accuracy of prediction. Our approach is similar to that of Aussem & Murtagh (1997) and is based on wavelet decomposition and predictions for each of the resolutions. The basic idea is to decompose the data into several resolutions, each reflecting a different frequency in the data. The data can then be expressed as an additive combination of the wavelet coefficients at the different resolution levels. Then, forecasting is done for each resolution level separately and the individual predictions are recombined to form the final forecast. Unlike Aussem & Murtagh, who used neural networks for prediction, we use simple linear models (Yu et al., 2001).

The wavelet transform is a synthesis of ideas from engineering, mathematics and physics. The Discrete Wavelet Transform (DWT) is similar to DFT or DCT in the sense that it decomposes the data into frequencies, but it has two important advantages over these methods: First, it quantifies location in time *and* frequency, i.e. it preserves information about both which frequencies exist in the data and in which time-intervals these frequencies appear (Polikar, 1996). Second, it is suitable for use with non-stationary data (Abramovich et al., 2000). These advantages are especially meaningful when designing an automated system.

To predict next day sales, we use the de-noised series. This series is decomposed into several resolutions. We then use a modification of DWT, called redundant (or stationary) wavelet transform, in which decimation is not carried out. For the purposes of the given paper we have implemented the algorithm as described in Aussem & Murtagh (1997) and Yu, et al (2001) (for theoretical and technical details see Appendix C). In our application we have five resolutions: the first four represent the high frequencies in the data (detail coefficients) and the last, the residual, reflects the low frequencies. Figure 8 illustrates the decomposition of the de-noised cough medication sales into five resolutions using the redundant wavelet transform. It can be seen that the higher resolutions capture the high frequencies in the data, while the lowest resolution captures the main trend.

Next, a one-step prediction is computed for each resolution separately using an autoregressive (AR) model. We use the AR model, which belongs to the category of linear models, for reasons of simplicity and interpretability (for information about AR models, refer to Appendix B).

In comparison to the complicated structure of the de-noised series, the wavelet coefficients within a certain resolution create a more regular series, in which case the AR models perform well. However, this layer of the detection system can be modified so that non-linear forecasting methods can be used.

After a new point is predicted for each resolution, the predictions are combined by summation to create the one-step ahead prediction of de-noised sales. Figure 9 illustrates the prediction for each of the resolutions separately, and their sum which is the prediction of the de-noised next day sales.

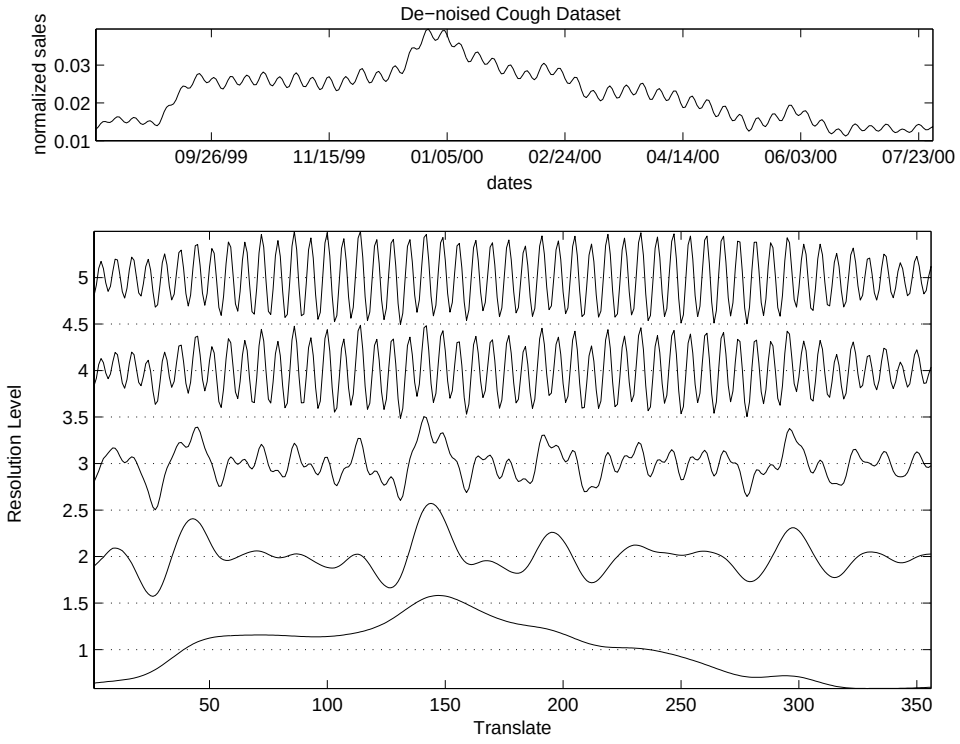

Figure 8: Decomposing the data into five resolutions using the redundant wavelet transform

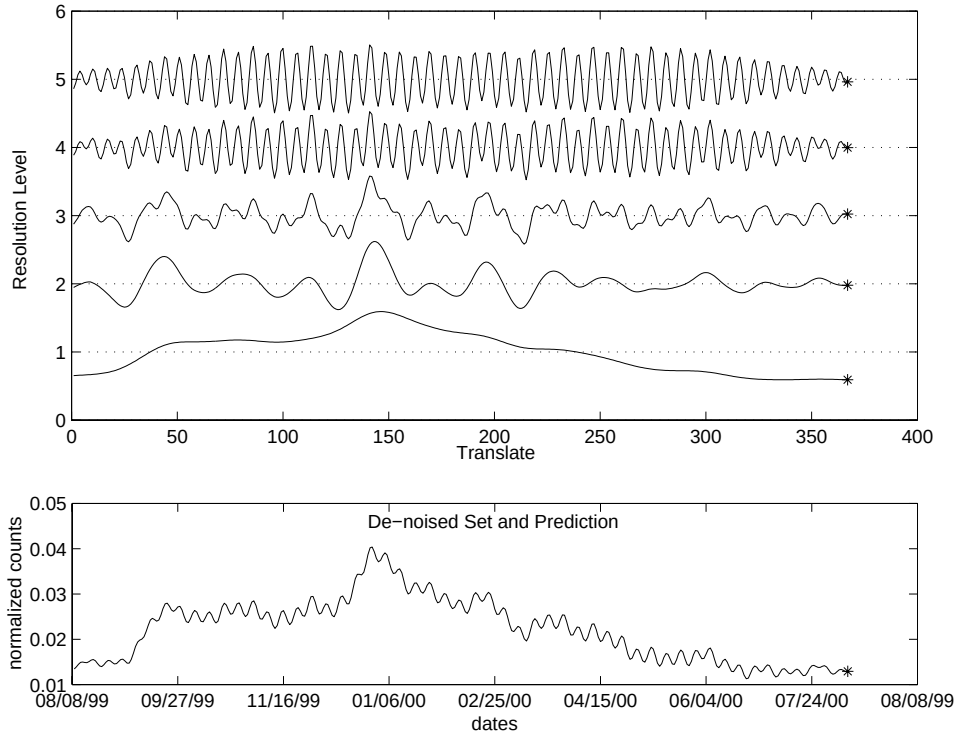

Figure 9: One step ahead predictions for each resolution and their sum

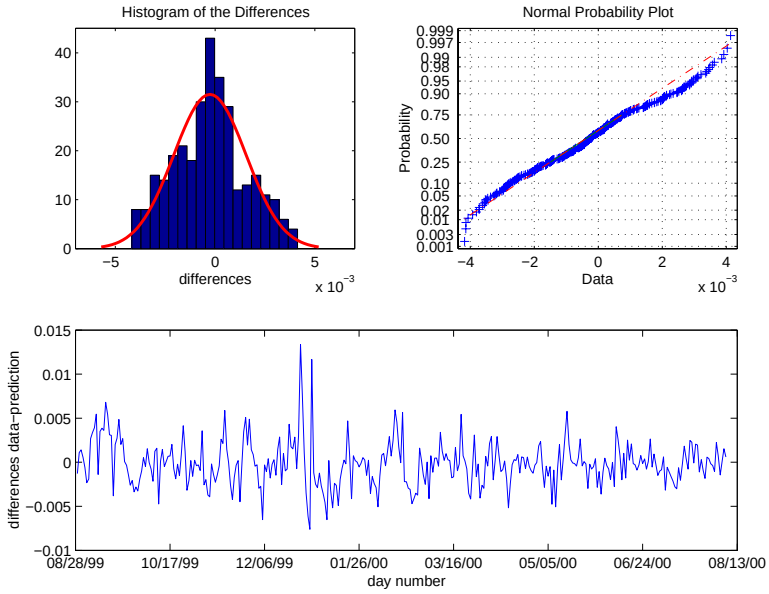

Figure 10: Checking the Normality Assumption of the differences between predictions and real data

3.4 Threshold Selection

Specifying the threshold is one of the critical decisions that must be made when designing a framework, since the threshold distinguishes between outliers and regular data flow. For example, by making the threshold closer to the prediction we increase the risk of a type I error, i.e. the risk of a "regular" (non-outlier) point falling beyond the threshold, but decreasing the type II error, i.e. the risk of not identifying a true outlier. In terms of the early detection system in relation to public health, the type I and type II errors correspond to taking preventive measures, i.e. spending money, on something that may not turn out to be an epidemic or an attack vs not identifying the problem on time and potentially resulting in loss of many lives respectively. Since public health and government money are at stake, it is not up to the system designer to set the threshold explicitly. It is to be assigned according to the government decision based on the tradeoff of losses described above.

For a given set of type I and II error probabilities we studied the distribution of the differences between the normalized observed data points and the corresponding predicted data points. The distribution of the differences was assumed to be approximately normal, since the differences are sums of various errors, and should therefore approximately follow a normal distribution. To check this assumption for specific data, we can use various statistical methods, such as histograms, probability plots, and more formal statistical tests. We illustrate how graphical methods can be used for assessing the normality assumption for the cough medication sales in Figure 10. From experiments on real data we learned that there might exist a very small percentage of differences that are extremely large (more than expected under a normal distribution). To account for this possibility we use a truncated version of a normal distribution fitting, where 10% of the empirical distribution tails are truncated. The threshold selection process is, in fact, very similar to boundary selection for Control Charts in Quality Control analysis (Montgomery, 1985). In control charts it is customary to select the boundary to be a multiple of the standard deviation of the distribution. It is also common practice to select three as this multiple. Since the true distribution is not known and it is not possible to compute the probability limits exactly, 3σ boundaries provide a good ap-

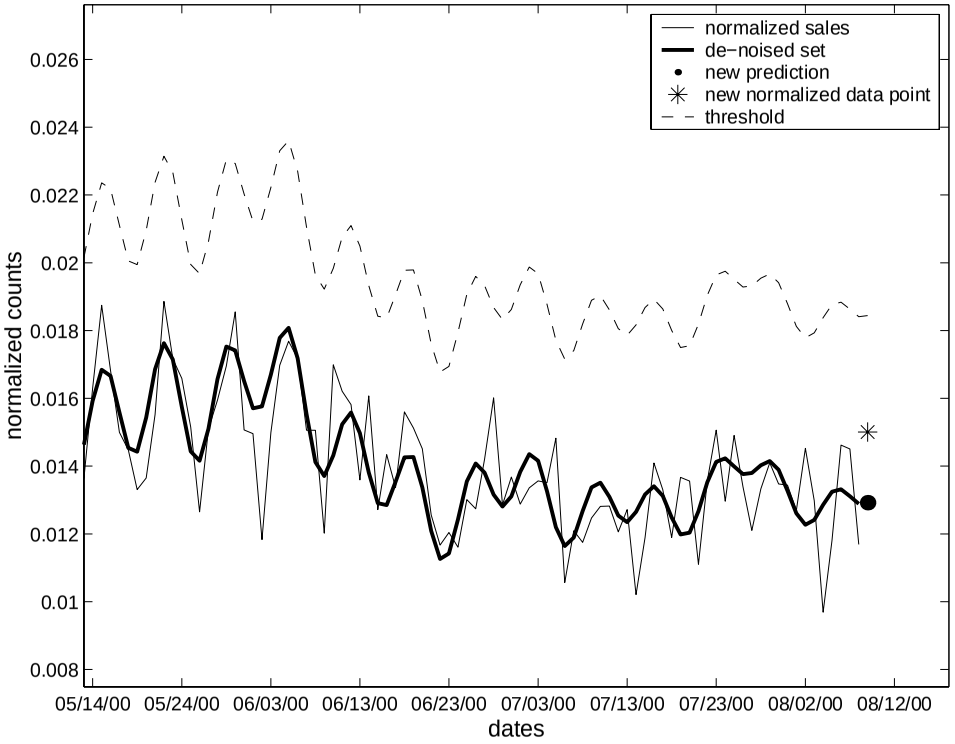

Figure 11: Comparing the next day prediction to the 3σ threshold for cough medication data

proximation to 0.001 boundaries if the normality assumption is valid and are also known to perform well in practice.

After the prediction is obtained and the next day sale information is available, the new data point is compared to the prediction+threshold value. If the data point exceeds the boundary, a flag is raised declaring the new data point an outlier.

Figure 11 illustrates the process of comparing a new data point to the 3-sigma threshold for the cough medication data. The new data point does not exceed the threshold, and is therefore considered an ordinary point.

The framework as described above was fully automated. The code for the project was written in Matlab and is available at http://cs.cmu.edu/~anya/BTproject/code/.

4 Evaluation

In an ideal situation, where the purpose of the project is to create an early detection system for a given bio-terrorist attack (e.g. anthrax), it is essential to be able to assign labels in the data denoting when and how the attack has occurred for proper evaluation. Unfortunately for the project and fortunately for the society, no data was available representing the course of an anthrax attack at that time. It is possible that grocery data from the three states where people were affected by anthrax was collected, however it was not available in time for the framework evaluation. Furthermore, the number of cases is very small (only two illnesses) for the framework used on grocery data to be effective.

Traditionally, there would be two major measures to evaluate a detection system: *sensitivity* and *specificity*. *Sensitivity* measures the rate at which the real alarms are detected and requires

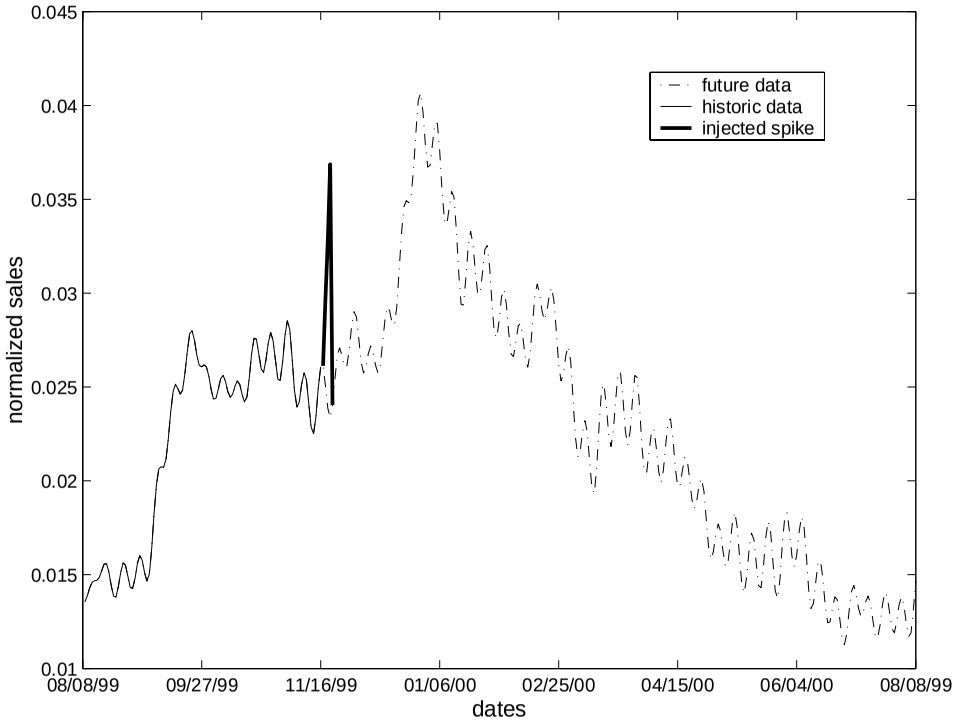

Figure 12: Hypothesized structure of the footprint of an anthrax outbreak in the sales data

the time and manifestation of an attack in the given data. *Specificity* measures the rate of the false alarms in the system and requires the knowledge about the absence of the abnormality in the data. It is used in epidemiology and medicine, which is the number of false alarms not flagged divided by the total number of data points that do not include a signal. Clearly, neither of the traditional measures are applicable in the given case, due to the lack of labeled data.

Farrington et al. (1996), who investigated an early detection system based on counts of voluntary reports of Salmonella, used an evaluation method based on simulation. Assuming that the counts are independent and distributed according to a Poisson or a Negative Binomial distribution, they generated random data from these distributions. Next, they added a hypothesized count at a certain point in the random data, and estimated the probability of detecting this added count. This procedure was repeated, adding a count of different magnitude each time.

4.1 "Spike"

We used a different method for evaluating our detection algorithm, which is more general and uses real data rather than simulated data.We turn the process into a supervised learning one, by injecting *spikes* that represent a simulated attack of a bio-agent such as anthrax. We consulted epidemiologists to obtain a valid simulation of an anthrax footprint. Figure 12 illustrates a possible footprint of an anthrax attack in sales of cough medication.

Anthrax is caused by a bacterium and if a sufficient amount of spores is inhaled it would causes fever, fatigue and difficulty of breathing. Death can occur within days. Based on this information, and with the assistance of Dr Mike Wagner from the Center for Biomedical Informatics at the University of Pittsburgh, we designed a pattern that spans three days with a linear increase in the rate of sales. Since, the spike can only be an approximation to the real case, the shape of the spike

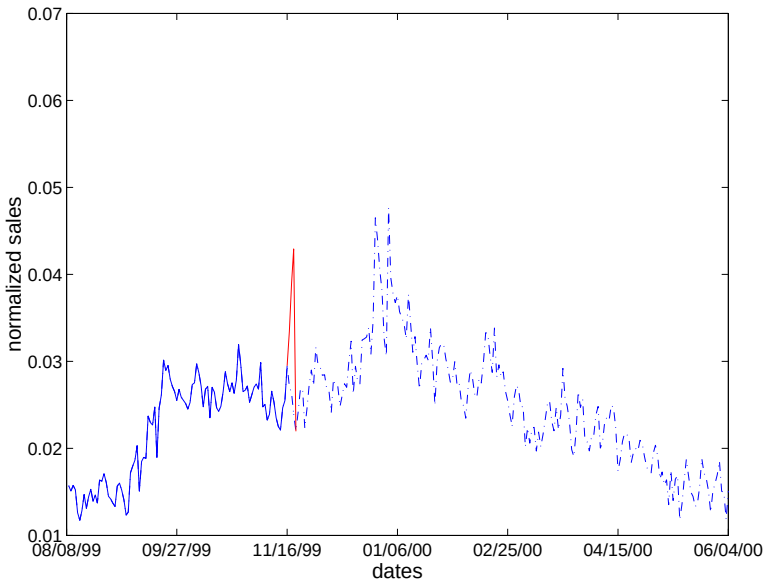

Figure 13: Adding the footprint into the daily sales data: the data with a 3-day pattern added on 11/16/99-11/18/99

is defined by two parameters: height and slope.

After the structure of the signal, denoted by $spike(height, slope)$,was specified, it was injected, i.e. added, at *every* point of the dataset (unlike the method by Farrington et al. where the spike was added in turn to one specific point in time). We considered all data points outside the range of the data that was "affected" by the spike to be "normal" . The addition of a simulated anthrax footprint in the cough medication data is illustrated in Figure 13. The figure is a snapshot of the addition of the three-day pattern on a specific day, resulting in a sharp increase in sales on that day and the following two days.

4.2 Spike Detection Ratio (SDR)

In order to test the detection rate of the framework, we applied the system, i.e. implemented framework, to each point of the data.set $spike(height, slope)$ was added. We define a measure that estimates the proportion of detecting spike of a specific pattern: the Spike Detection Ratio (SDR). This is the proportion of spikes detected divided by the total spikes added.

$$SDR_i = \left(\frac{\text{spikes detected}}{\text{spikes injected}} \right)_i \tag{6}$$

, where $i = 1, 2, 3$ stands for detection within the i next days.

For a dataset with n points (excluding the first points that are used for initializing the forecasting component) and a pattern that spans k time points (e.g., days), the SDR denominator equals $n - k$. This statement is valid under the assumption that the pattern is added at every time point, excluding the initialization period.

To calculate the false alarm rate, we add a pattern with spikes of height zero and count the number of detections. This is technically equivalent to looking for detections without spikes added at all. The information about how many spikes were detected can then be used to compare the performance of various configurations of the system.

By applying such an analysis it is possible to evaluate different aspects of the data and the framework. Firstly, it is possible to reveal the characteristics of the data that is most sensitive to abnormality occurrences. It is also possible to perceive the type of spikes that are easier to detect with the given system. Sensitivity of the system can be tested based on how early the spike was detected. Secondly, it enables the identification of different types of footprints that are easier to detect with a given system than others (determined, for example, by the height, and slope of a pattern). This can be used further so that the type of pattern detected can imply the type of outbreak and its associated symptoms.

5 Experiments

Data used in the project is the daily retail over-the-counter medication sales. The information is obtained by summation of sales for individual products over the period of a single day. There were 541 points available starting August 8, 1999 until January 31, 2001 for several different product categories such as Cough Syrup/Liquid Decongestant; Tabs and Caps (include Advil Cold/Sinus, Tylenol Flu Non-Drowsy, etc); Throat Lozenges/Cough Drop and Nasal Spray/Drops Inhalers (for morede detailed data description, refer to Section 2).

We illustrate experiments carried out using the cough dataset on Figure 14 (top graph). Similar experiments were carried out on other datasets. Results for these series are reported in Section 6.

First, the pre-processing step was applied. In the pre-processing step the cough medication dataset was normalized by the total sales of health category. Datapoints around holidays on which the stores were closed were interpolated as described in Section 3.1, resulting in the dataset as appears in Figure 14 (bottom grah).

It is evident from Figure 14 that the normalization and zero-interpolation steps have achieved the desired effect. For example, original data dips around January 1, 2000, turn into increases on that date after pre-processing. This means that even though there were less purchases of cough medications overall, the sales volume of cough medications on that date was still higher than expected. This important information would not be revealed if the pre-processing step was absent.

After the pre-processing, the dataset was divided into subsets and the simulation of the framework usage was carried out. Twenty-nine data points were reserved for the first model selection and prediction of the 30th data point. The $spike(height, slope)$ was added to points 30, 31, 32. The algorithm (as described in Section 3) was applied and the 30th point with an injected spike was compared to the threshold. Then, the first day detection measure SDR_1 of the first spike was recorded. Next, the framework as described was applied to points 31 and 32 to record the first spike detection within the first two and three days. The same spike was then added to the normalized dataset starting at point 31. The process was repeated, rolling forward one data point at a time, consequently obtaining three sets of 509 points of detection for evaluation.

Spikes of various shapes were tested. The heights of spikes tested were:

$$\{2 \times (\text{data range}), 2 \times (data\ range) - \frac{2\times(data\ range)}{4},$$
$$2 \times (data\ range) - \frac{2\times(data\ range)}{4} - \frac{2\times(data\ range) - \frac{2\times(data\ range)}{4}}{4}, \cdots, \frac{data\ range}{2^{24}}\}$$

The total of 26 different spike heights were tested. More spikes of smaller heights were tested, as the system performed well on higher spikes. The system seemed to have stabilized at the bounds. If the height of the spike was too small, such as $height = \frac{(data\ range)}{2^{24}}$, it had approximately zero probability of being detected. If the spike was too high, such as $height = 2(data\ range)$, it had

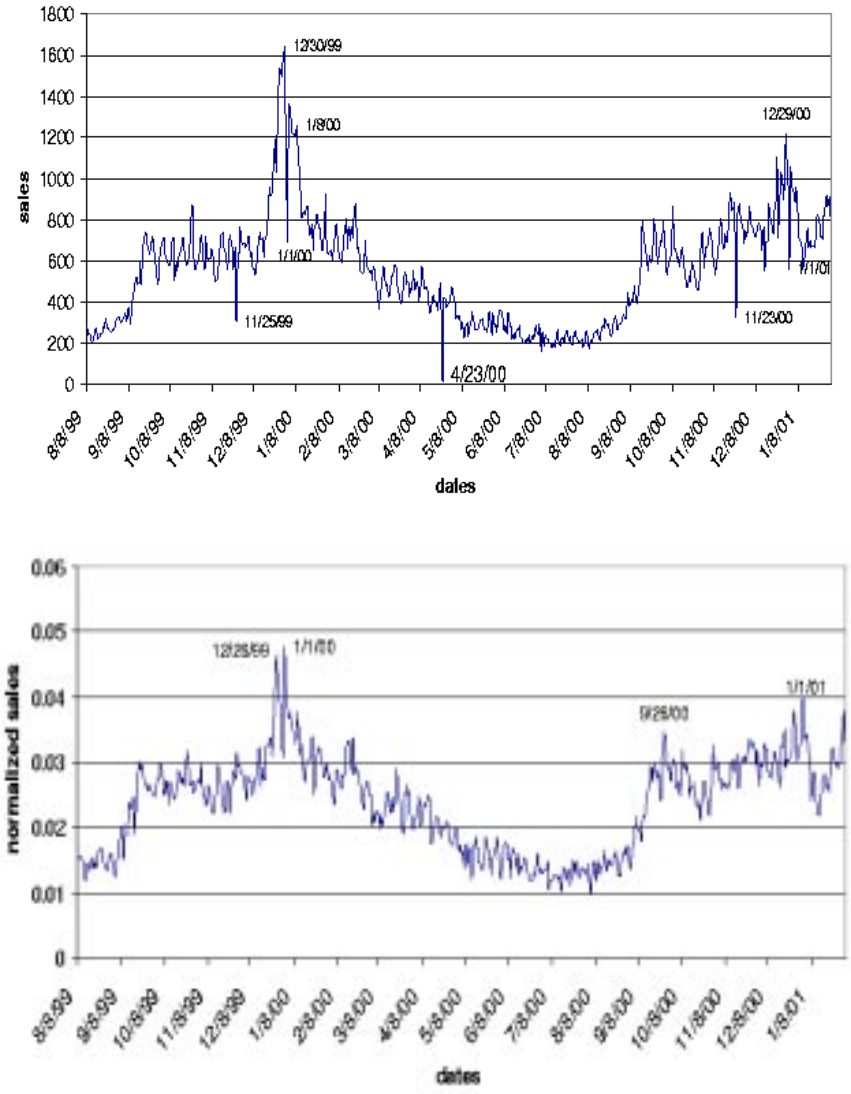

Figure 14: Cough dataset prior to(top) and after (bottom) pre-processing

100% chance of being detected.

The slopes tested were $\{0, \frac{1}{7}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2}, 1\}$, where 0 slope was used for identification of false alarms, slope of $\frac{1}{7}$ means that the medication sales were steadily increasing over 7 days. A slope of 1 represents a square wave. Please note, that only detection on the first day, first two and first three days was studied. Such constraint is based on the assumption that if anthrax was not identified within the first three days, it would have been too late.

Several configurations of the proposed framework were compared. In the following naming convention *D* stands for system configuration where prediction is done directly from normalized counts, i.e. the de-noising (DCT) step is not invoked; *M* stands for a configuration where the de-noising step is utilized prior to prediction; *AR1* and *AR7* stand for auto-regressive predictive models AR(1) and AR(7) respectively; *W* means that the configuration involved invoking wavelet decompostion prior to applying forecasting models. Below are the configurations that were tested

DAR1 — the simplest system with a predictive AR(1) model used to make predictions directly from the normalized data

DAR7 — a predictive model AR(7) was used to make predictions directly from the normalized data

MAR1 — the system includes a de-noising step and AR(1) for prediction

MAR7 — the system includes a de-noising step and AR(7) for prediction

MWAR1 — all layers of the framework utilized, with de-noising, Wavelet decomposition (5 wavelet components) and AR(1) predictions

MWAR7 — most complicated system. All layers are utilized as in MWAR1. Except the predictions were obtained using AR(7).

AR(1) was selected as a representation of the simplest predictive model. AR(7) was utilized as a more complicated predictive model intending to capture the seven day periodicity effect.

6 Results

One possible anthrax scenario can be represented by a spike with $slope = \frac{1}{3}$. Figure 15 illustrates Spike Detection Ratios (SDRs) for different heights detected on the first, within the first two and within first three days for the framework as defined in Section 3.

The x-axis of Figure 15 represents the spikes' height. A spike ratio of 1, $SDR = 1$ means 100% detection of injected spikes anywhere in the data. It can be noticed that the lower the height of the spike the harder it is to detect. However, if a spike reaches the height of $\frac{19}{14} \times$ (data range), it is detected with the probability of 1 within three days in the Cough medications subgroup sales data. It is also evident from the graph that the rate of detection increases on the second and third days, meaning that even if the spikes were not detected on the first day they still have a chance of being detected within the first two or within three days.

6.1 System Comparison Using The Cough Dataset

To compare several systems numerically we approximate the area under each curve displayed on Figure 15 for respective detection days ($i = 1, 2, 3$). We introduce a measure of a system's performance:

$$SSDR_i = \sum_{\forall heights} SDR_i \tag{7}$$

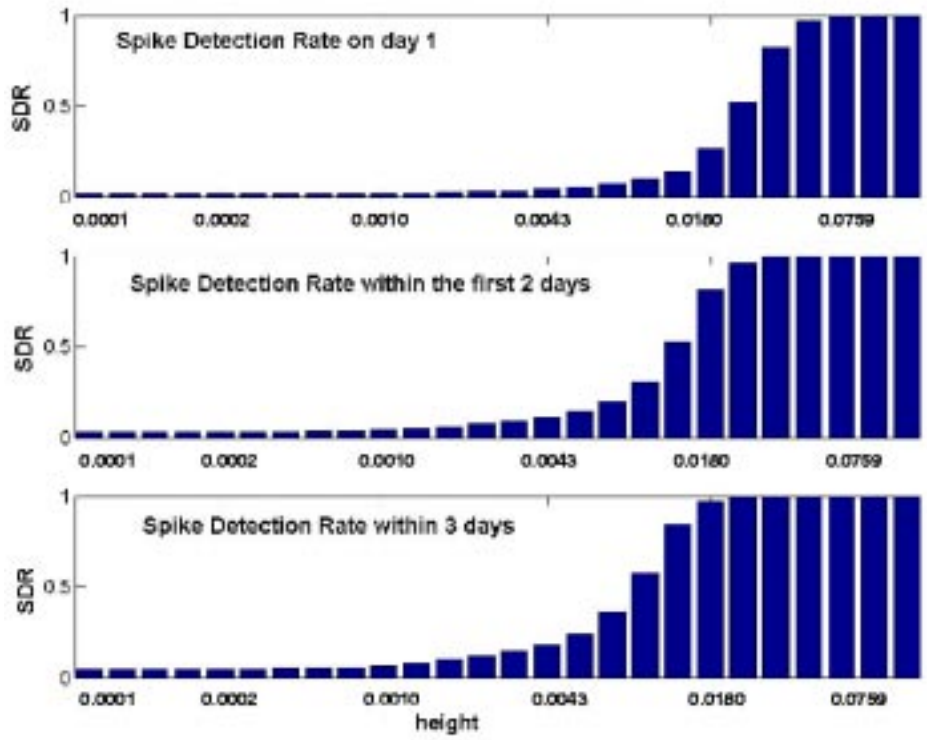

Figure 15: The Spike Detection Ratio as a function of the pattern's height, within a single day, two days, and three days (spike $slope = \frac{1}{3}$)

$COUGH$	$1^{st}day$	$2^{nd}day$	$3^{rd}day$
DAR1	6.2480	7.4429	8.2657
DAR2	5.9567	7.3661	8.1929
DAR3	6.1398	7.6102	8.6063
DAR4	6.0886	7.6043	8.6398
DAR5	6.0472	7.6260	8.7323
DAR6	6.0315	7.6654	8.7323
DAR7	6.0846	7.7126	8.7736
DWAR1	6.2382	8.1004	9.1890
DWAR3	6.1220	7.6693	8.7815
MAR1	6.3248	8.5453	9.8701
MAR3	6.2067	8.3209	9.6220
MAR7	6.3957	8.6417	9.9843
MWAR1	6.3209	8.6299	10.0630
MWAR3	6.2264	8.3681	9.6732
MWAR7	6.3760	8.6142	9.9606

Table 1: Comparing the outlier detection systems for the cough product category

which is an approximation to the area defined by $\int_{\forall heights} SDR_i$. This results in three values for a given spike slope.Several systems were compared based on the $SSDR$ measure and the slope of $\frac{1}{3}$. The results are in Table 1.

A slightly larger selection of models were applied to the Cough/Syrups product category. $DAR2, \cdots, DAR6$ are named using the same convention as DAR1 and DAR7 as described above, i.e. the systems were using predictive models $AR(2), \cdots, AR(6)$ respectively to make predictions directly from scaled counts. DWAR1 and DWAR3 denote systems that used wavelet decomposition, where AR(1) and AR(3) respectively were used to make predictions for each of the wavelet resolutions. These methods were later eliminated from the set of systems used for other products, since their performance was not as good as the other systems.

From Table 1 we can see that MAR7 was able to detect the largest number of outliers on the first day and within the first two days but MWAR1 performed the best within the first three days. In terms of the problem, it means that MAR7 was better at detecting the outbreak sooner whereas MWAR1 was better at detecting hard outbreaks. In fact, when analyzing the performance of the two systems at different heights it was evident that all three systems identified the outbreaks equally well when $height <= \frac{1}{4} \times$ (data range) and though the systems performed similarly well, MWAR1 has performed better with smaller heights. However, when the heights were in the interval $[0.0008, 0.024]$, MAR7 outperformed the more complicated system MWAR1, meaning that increasing the complexity of the system in this case did not add any additional value to the system's performance. The analysis is also supported by a graphical representation of the SDR-based system evaluation on Figure 16.

6.2 False Alarm Rates for Cough Dataset

The false alarm rate estimated using the SDR measure for a spike of slope 0 is recorded in Table 2 for selected systems that were applied to Cough dataset.

The numbers in Table 2 are obtained by injecting spike of zero slope and then performing the

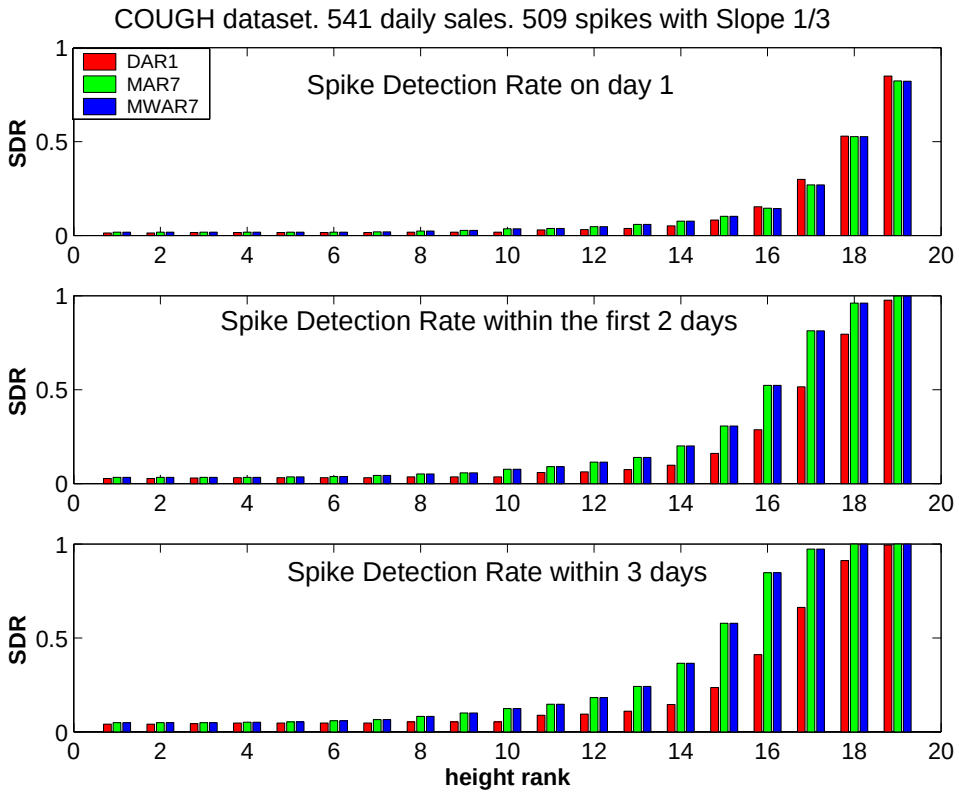

Figure 16: Comparing three configurations of the detection system using SDR-based evaluation

System	False Alarm Rate
DAR1	0.041
MAR1	0.043
MAR7	0.043
MWAR1	0.043
MWAR7	0.043

Table 2: False Alarm Rates for a variety of systems tested on the Cough dataset

100%	Height (*range)		
$System^{Slope}$	1/7	1/3	1
DAR1	2	$\frac{21}{25}$	$\frac{12}{25}$
MAR7	$\frac{9}{8}$	$\frac{12}{25}$	$\frac{12}{25}$
MWAR7	$\frac{9}{8}$	$\frac{12}{25}$	$\frac{12}{25}$

50%	Height (*range)		
$System^{Slope}$	1/7	1/3	1
DAR1	$\frac{1}{15}$	$\frac{1}{37}$	$\frac{1}{50}$
MAR7	$\frac{1}{50}$	$\frac{1}{118}$	$\frac{1}{156}$
MWAR7	$\frac{1}{50}$	$\frac{1}{118}$	$\frac{1}{156}$

Table 3: Spike parameter bounds for 100% detection (on the left) and 50% detection (on the right).

evaluation as described in Section 4. If the system detects an outlier in the dataset where the zero slope spike is injected, it is treated as false alarm. The closer the SDR value is to zero the better the performance is. Table 2 shows that the simplest system had the best performance of the systems compared. However, that could be attributed to the fact that the simplest system is poor for outlier detection in general. Since systems $MAR1, MAR7, MWAR1, MWAAR7$ had performed identically it means that there was some noise in the data that was not captured in the de-noising process and that due to its simplicity, DAR1 was unable to capture. It is important to note, that increased complexity of the system does not negatively influence system's performance.

6.3 Bounds for Spike Parameters

When analyzing early detection systems based on simulated data it is important to identify what kind of spikes could be detected. In order to estimate bounds for spikes' parameters, i.e. heights and slopes, that the proposed framework would detect best, we selected the smallest height and slope values such that 100% of spikes were detected for the upper bound. The lower bound was found based on the 50% detection ratio. These bounds identify spike shapes that could be detected anywhere in the dataset with probability of 1 (on the left) and probability of 0.5 on the right. The bounds for parameters, i.e. the largest and smallest values for height and slope, for three different configurations of the framework, namely DAR1, MAR7 and MWAR7 are presented in Table 3.

Spike parameters shown in Table 3 also support the conclusion that systems MAR7 and MWAR7 have similar performance, meaning that they are able to detect spikes with similar probability. Depending on the spike slope, there may be significant improvement of more complicated systems over the simpler ones. Table 3 shows that if the spike rises gradually, for example $slope = \frac{1}{7}$, complex systems such as MAR7 and MWAR7 outperform DAR1 significantly.

6.4 Dataset Study

Another important aspect of early detection system evaluation is to be able to draw conclusions about the suitability of the system for the particular data. For this purpose a *spectral* graph was introduced. This graph allows to make judgements about system's performance at a particular point in the data. The *spectral* graph applied to the framework as described in Section 3 to the Cough dataset is shown on Figure 17.

Figure 17 should be interpreted as follows: The x-axis of the graph is the time axis, the same as the one used for displaying the data. The y-axis indicates the height of the spike injected. For each point in time, there is a gray-scale bar that is present if the spike of a given height was detected using the described system. Spikes of $slope = \frac{1}{3}$ were injected. The bars range from darkest, representing the highest spike injected, to the lightest shades of gray, representing spikes of the lowest height. At each time point, the higher the bar the more spikes were detected. For example, it is evident from the graph that there are several high bars and the rest of them are smaller. The highest bars appear on days like January 1, 2000 when the sales were on the rise and had much

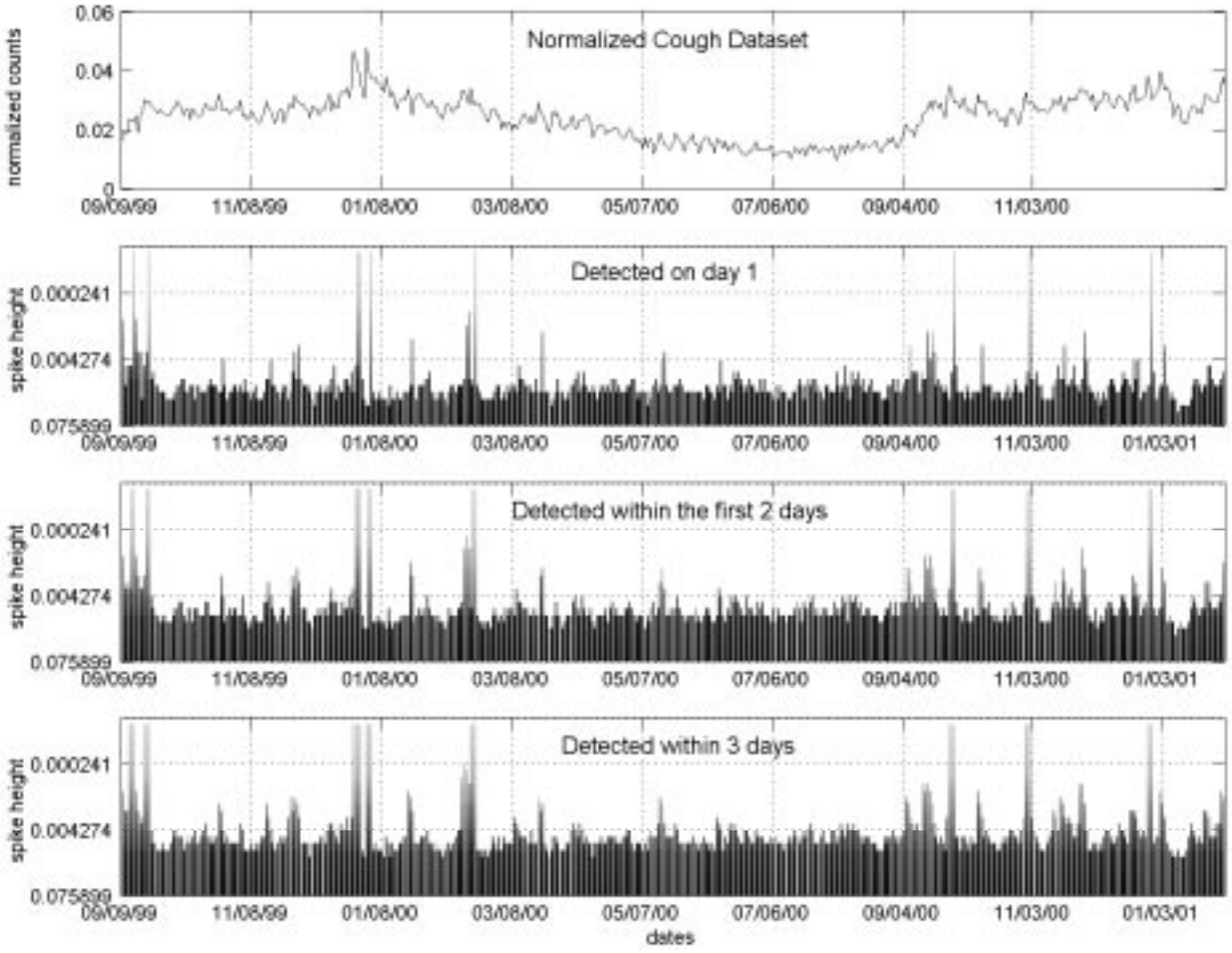

Figure 17: Spectral Graph for Cough Dataset. The height of each bar represents the proportion of spikes detected for different heights. Shade of gray represents the height of the injected spike (darkest=highest). $Slope = \frac{1}{3}$

$TABS\&CAPS$	$1^{st}day$	$2^{nd}day$	$3^{rd}day$
DAR1	5.5453	6.9390	7.8780
MAR1	5.2677	7.2421	8.6339
MAR7	5.5591	7.8878	9.3465
MWAR1	5.4862	7.8169	9.2559
MWAR7	5.5453	7.8701	9.3661

Table 4: Comparing the outlier detection systems for the tabs and caps product category

$THROAT$	$1^{st}day$	$2^{nd}day$	$3^{rd}day$
DAR1	5.5453	6.9390	7.8780
MAR1	5.2677	7.2421	8.6339
MAR7	5.5591	7.8878	9.3465
MWAR1	5.4862	7.8169	9.2559
MWAR7	5.5453	7.8701	9.3661

Table 5: Comparing the outlier detection systems for the throat lozenges product category

higher values than on other days. The same can be observed with the rest of the high bars, i.e. more spikes were detected during the holidays. However, it was also noticed that the false alarms were detected on the holidays as well, meaning that the holidays in general are very sensitive to being detected and perhaps a more sophisticated analysis has to be applied to study the holidays separately or in addition to the routine analysis. It can be noticed however, that the system is good at detecting spikes when the series is on the rise, not only during holiday season.

6.5 Other products

We tested our framework in various configurations on other product categories. It performed similarly to its performance for the cough dataset. We, therefore, omit the detailed analysis for those categories. We present comparison of the different configurations in Tables 4, 5, 6. We compare performance of our detection system when applied to Tabs&Caps, Throat and Nasal subgroups are using SSDR measure (refer to Equation 7).

For the Tabs&Caps product category data, MAR1 outperformed the other systems when comparing detection on the first day and within the first two days. MAR7 and MWAR7 performed equally well for detection within the first three days. MAR1 outperformed MAR7 and MWAR7 on first day and within three days detection, and thus had the best overall performance.

Table 5 shows that for the throat product category MAR1 seemed to have the highest detection rate on the first day, however MWAR1 showed superior performance detecting outliers within the first two and three days.

The nasal inhalers product category seemed to have the smallest rate of detection compared to the four subgroups of products considered. Such results can be expected based on the low variance of the data which can be seen on Figure 1. Also, DAR1 outperformed other systems in the first day detection. However MWAR7, the most complex system, seemed to detect most outliers during the first two and three days.

$NASAL$	$1^{st}day$	$2^{nd}day$	$3^{rd}day$
DAR1	5.5453	6.9390	7.8780
MAR1	5.2677	7.2421	8.6339
MAR7	5.5591	7.8878	9.3465
MWAR1	5.4862	7.8169	9.2559
MWAR7	5.5453	7.8701	9.3661

Table 6: Comparing the outlier detection systems for the nasal inhalers product category

6.6 Discussion

From the experiments we found that MAR1 detected most outliers across all products on the first day; system MWAR1 detected most outliers during the first two days; and system MWAR7 had the best performance in outlier detection within three days. Such results can be attributed to the fact that to detect a sudden or a gradual rise in the data it is best to look at the least number of previous data points possible. If the model takes into account more than one previous day it is less likely to detect the gradual rise.

It is also clear that the introduction of the DCT decomposition step improves performance. However, it is not at all obvious that the wavelet decomposition step prior to the prediction makes a difference. It seems that the more complex systems, i.e. MWAR1 and MWAR7, are better at detecting "hard-to-detect" outliers. They detect outliers on the third day when other systems failed to detect those outliers completely.

There are certain results that were expected and were supported experimentally: higher spikes with steeper slopes are detected better than spikes with slower rise and lower height. Also, if there is a naturally occurring rise in the data, the spike that is injected at that point has a higher probability of being detected. That, however, leads to detection of holidays at all times. In fact, most of the false alarms were due to holidays. It is expected, though, that as more data is collected, the more stable the system would become adapting to re-occurring holidays.

When considering systems in the context of the problem, it is not necessary to raise a flag for an epidemic if the irregularity in the data is not too strong, but it is definitely important to call the attention of the people monitoring the system when the data exhibits abnormal behavior. The life of many people may be at stake and therefore it is important that the system be sensitive enough to signal a potential disaster. Perhaps, it would be useful to use a combination of several approaches. For example, two predictive methods could be applied simultaneously. If only one of the systems triggers, the signal should serve as a *yellow* warning flag, rather than the *red* outlier flag.

From testing different systems on a set of products, it becomes obvious that certain systems are better for a specific type of products. Simpler system might work on data that is less irregular while more complicated ones are required when a simple model cannot be used to describe the data. Results of this research suggest that when a new product is being tested for abnormalities, several systems should be applied to identify the one(s) that suit the particular product category best.

Early detection of epidemics and bio-terrorism attacks is a problem that requires further research not only in the areas of surveillance system identification but also in the field of epidemiology and marketing science. The usage of non-symptom specific data provides an opportunity to warn the population about the possibility of the disaster and perhaps save more lives than is now possible using the existing methods. However, it is not at all clear how to determine the manifestation of

an outbreak in retail data. The spike injection evaluation technique is an attempt to analyze the data as if information about such manifestation was known.

7 System Limitations

The framework proposed in this research is an attempt to create an automated comprehensive early detection system based on non symptom-specific data. We hope that this system will serve as a basis for further developments in the area. It is important, therefore, to acknowledge the framework's limitations.

The dataset available for testing is not extensive enough to be able to trace all the periodicities and trends intrinsic to the data. For example, there is a clear abnormality in purchasing behavior around holidays, however the data only spans one set of holidays. This means that there is only one data point for the system to learn about an important feature of the data. Obviously, it is hard to evaluate the behavior of the system if it had seen enough data to make an adjustment for such trends.

The framework proposed is a complicated system that is a combination of several techniques and composed of several different layers. Each step involves estimation of at least one or even several parameters. For example, it is clear from our experiments that the DCT step adds significant improvement over predictions made directly from the scaled data, but a careful DCT coefficient selection procedure is a complex and time-consuming task. The number of wavelet coefficients is another free parameter. It is easy for a human eye to detect that the fourth component after the transformation is smooth enough to stop the decomposition, but it is not easy to automate this selection.

8 Summary

The problem of early detection of epidemics and bio-terrorism attacks has been researched for centuries. There are surveillance systems in place that evaluate epidemiological data received from hospitals and laboratories. However, these systems cannot detect as *early* as it is desired. Initiatives have been taken to research the issue and steps are being taken in many directions in order to create an *early detection* system. Preliminary experiments conducted give strong evidence that non symptom-specific based systems that were overlooked before, may indeed be accurate enough to have a right to be considered. Complete framework was developed that can serve as a prototype for such system. It is flexible enough to incorporate more advanced prediction techniques as they become available and to be applicable to other datasets that have time series properties and satisfy the condition of being correlated with bio-terrorism attack that is to be detected. The system was applied to several different medication categories and encouraging results were obtained. The results support the claim that non symptom-specific data can improve timeliness, sensitivity and overall performance of detecting and outbreak over the existing early detection systems. The automated framework described is a strong basis for future research.

9 Future Work

As mentioned before, even though the presented framework for an early detection system is extensive, there is room for improvement. Even though most of the parameter selection steps are automated, it is desired that there are no points of arbitrary parameter settings.

The skeleton of the system is solid, but the methods applied in each step are interchangeable. The linear models were considered in the current configuration to obtain better understanding about the system, but it is possible that more complicated methods like time delay neural network (TDNN) can improve the performance by providing more accurate predictions. Preliminary experiments did not show any significant improvement, though. A more exhaustive search of non-linear systems would be highly beneficial.

An important way to justify the number of outliers that are detected/missed by the system is to conduct a thorough cost sensitivity analysis, which would determine the threshold selection. In fact, it is possible that such cost sensitivity analysis would also be useful in de-noising step when applied in a retrospective manner.

Another aspect of the system needs to be explored further as well. Suppose that an attack has been carried out and then successfully detected. The post-detection knowledge about existing outliers needs to be incorporated into the system. It would also be beneficial to take into account information about public events, such as football games or floating holidays. Such expert information could have significant impact on the system's performance.

Since only 541 data points were available for training the system, the overhead of using all available data for building models, selecting thresholds and predicting new day sales, was not too great. To achieve best accuracy it is usually preferable to use as much information as possible, however if the system were implemented in the real world and much more data has become available in a short period of time or there was an intention to use the system for the more refined scale data, such as hourly or even as refined as individual purchase records, it is essential to look into online learning algorithms for more efficient performance.

It is suggested to look at other details available about the purchases made at grocery stores such as the location of the purchase, customer information, etc. If, for example, it were known that an increasing number of adults ages $25 - 30$ with no prior record of re-occurring medication purchases have suddenly significantly changed their behavior it could perhaps be a good indicator of an epidemic or even a bio-terrorism attack.

Another idea that might require cooperation of many areas of research related to surveillance systems is the incorporation of systems that rely on non-symptom specific data with the more traditional systems based on medical information. These systems, though not always timely, have proven to be stable and reliable as a source of problem identification information.

10 Acknowledgments

I would like to thank Rich Caruana, Galit Shmueli and Sebastian Thrun for their invaluable help with participating in brainstorming sessions and assisting to make the project stronger. I would also like to thank David Deerfield and Laura F. McGinnis for their help with the dataset and Mike Wagner for his help with gaining more understanding about the epidemiological data and spike simulation.

A Discrete Cosine Transform (DCT)

DCT is closely related to Discrete Fourier Transform (DFT). In fact, the only major difference is that the cosine transform approximates the function using cosines and not complex cosine-sine functions, working on the real scale. Original data is represented by vector $[x_1, x_2, \cdots, x_N]$. The result of the DCT transform is a vector of numbers $[y_1, y_2, \cdots, y_N]$ calculated according to the

following formula:

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos \frac{\pi(2n-1)(k-1)}{2N}, \ k = 1, \cdots, N \tag{8}$$

, where $w(k) = \begin{cases} \sqrt{1/n} & , k = 1 \\ \sqrt{2/n} & , 2 \leq k \leq N \end{cases}$ To reconstruct the original series, the inverse discrete cosine transform (IDCT) is used, mathematically described as follows:

$$x(n) = \sum_{k=1}^{N} w(k) y(k) \cos \frac{\pi(2n-1)(k-1)}{2N}, \ n = 1, \cdots, N \tag{9}$$

and $w(k)$ is defined as above. For further details on DCT and IDCT, refer to Jain (1989) and Pennebaker et al. (1993).

B Predictive Linear Models: Autoregressive Models.

An AR model describes the series at time t as a weighted average of observations at previous times. An autoregressive model of order p can be described by the following equation:

$$x_t = \sum_{i=1}^{p} a_p x_{t-p} + \epsilon_t \quad t = 1, 2, \ldots \tag{10}$$

where x_t is the value of the series at time t; ϵ_t is White Noise with zero expectation and a constant variance (WN(0,σ)); x_t is uncorrelated with ϵ_t; and the coefficients satisfy $|a_i| < 1$. The coefficients can be estimated from the data by the method of maximum likelihood. For further details see Hamilton (1994).

C Wavelets

The idea of wavelet analysis is to decompose a series using a set of functions, called wavelets, that are suitable for capturing the local behavior of non-stationary series (Shumway & Stoffer, 2000). As Fourier analysis uses sine and cosine functions for the decomposition, wavelet analysis uses the "father" and "mother" wavelet functions, denoted by ϕ and ψ. These functions are used to capture the low- and high-frequency components of the data, respectively. More wavelet functions are then generated from the father and mother wavelets by an operator called *translation* and by refining the resolution (or *scaling*):

$$\phi_{j,k}(t) = 2^{-j/2} \phi \left(\frac{t - 2^j k}{2^j} \right) \tag{11}$$

$$\psi_{j,k}(t) = 2^{-j/2} \phi \left(\frac{t - 2^j k}{2^j} \right) \tag{12}$$

where $2^j k$ is the translation parameter and 2^j is the resolution parameter. This means that as the resolution increases, the wavelet function becomes more spread out and shorter (Shumway &

Stofer, 2000). This is equivalent to passing the data once through a low-pass filter and applying a binary decimation operator, and once through a high-pass filter followed by binary decimation. This procedure is then re-applied to the resulting two series, and continues on recursively (Nason & Silverman, 1995).

The Discrete Wavelete Transformation (DWT) results in two sets of coefficients: smooth coefficients ($s_{j,k}$) that represent the smooth behavior or low frequencies in the data, and detail coefficients ($d_{j,k}$) representing the high frequencies in the data.

The redundant (or stationary) discrete wavelet transform (RDWT) differs from DWT only by the decimation step. The high and low pass filters are applied to the data at each level to produce two sequences at the next level, but without carrying out the binary decimation. This means that the resulting sequences each have the same length as the original sequences (Nason & Silverman, 1995).

A fast implementation and computation of RDWT was suggested by Aussem & Murtagh (1997) and further extended by Yu, et al (2001).

Let H be a low pass filter, for example a B_3 spline defined as $(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16})$ and $r_0(t) = x(t)$, where $x(t)$ is the data vector. Then wavelet coefficients can be computed according to the following equation:

$$r_i(t) = \sum_{t=-\infty}^{\infty} h(l) r_{i-1}(t + 2^{i-1} l),\ i = 1..p \tag{13}$$

, where p is the number of wavelet resolutions.

The high level resolutions are then found by:

$$d_i(t) = r_{i-1}(t) - r_i(t) \tag{14}$$

The last resolution denoted by r_p is the low resolution component sometimes referred to as residual. Hence, set $\{d_1, d_2, \cdots, d_p, r_p\}$ represents the wavelet transform of the data up to resolution level p (Yu, et al, 2001). The inverse transform is then found by:

$$x(t) = \sum_{i=1}^{p} d_i(t) + r_p(t) \tag{15}$$

From Formula 13 it is evident that some guess for the values into the future needs to be provided. The effect is commonly known as boundary handling (Aussem & Murtagh, 1997) or boundary treatment (Yu, et al, 2001). There are different ways of handling the boundary. Some common ways are reflective or periodic boundary treatments as mentioned in (Aussem & Murtagh, 1997), however Yu, et al (2001) showed that frequently used ad-hoc boundary treatments lack in performance as they may not capture characteristics of the dataset at hand, hence they have proposed Periodic Boundary Treatment as described in Yu, et al (2001).

The wavelet algorithm for the specific project was implemented as described in Yu, et al (2001).

References

[1] Abramovich F., Bailey T. & Sapatinas T. (2000),"Wavelet analysis and its statistical applications", *The Statistician - Journal of the Royal Statistical Society, Ser. D*, vol. 49, pp. 1-29.

[2] Airhart, M. (1997). Why do we get the flu most often in the winter? Are viruses more virulent in cold weather? Scientific American, Ask Experts section.

[3] Aussem A. and Murtagh, F. (1997) Combining Neural Network Forecasts on Wavelet-Transformed Time Series. Connection Science, Vol 9. No. 1, 113-121.

[4] Boatwright P., McCulloch R., and Rossi P. (1999). Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model. (315-337).

[5] Brigham, E. (1988), The Fast Fourier Transform and Its Applications, Englewood Cliffs, NJ: Prentice-Hall, Inc.

[6] Brockwell, P. and Davis, R. (1996) Introduction to Time Series and Forecasting Springer.

[7] Chatfield, C. (1998) What is the best Method in Forecasting? Journal of Applied Statistics. 15:(19-38).

[8] Daubechies, I. (1992). Ten Lectures on Wavelets CBMS-NSF Series in Applied Mathematics 61, Philadelphia: SIAM

[9] Farrington C.P., Andrews, N.J., Beale, A.D., and Catchpole, M.A. (1996) "A Statistical Algorithm for the Early Detection of Outbreeaks of Infectious Disease" , *Journal of the Royal Statistical Society A*, vol. 159, part 3, pp. 547-563.

[10] Goldman L. (2000). Inhalation Anthrax in Cecil Textbook of Medicine. New York: Saunders Company.

[11] Graps, A. (1995). An Introduction to Wavelets. IEEE Computational Science and Engineering. Los Alamitos, CA: IEEE Computer Society.

[12] Hamilton J., (1994) Time Series Analysis Princeton University Press, Princeton, New Jersey.

[13] Jain, A. K. (1989), *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice-Hall.

[14] Mitchell, T. M. (1997). Machine Learning McGraw-Hill, Boston, Massachusetts.

[15] Montgomery, A. (1997). Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data.

[16] Montgomery D., (1985) Introduction to Statistical Quality Control. New York, John Wiley & Sons.

[17] Nijs, V., Deimpe, M., Steenkamp JB. and Hanssens, D. (2001). The Category Demand Effects of Price Promotions Under Review at Marketing Science, second round.

[18] Peña, D., Tiao, G. C., and Tsay, R. S. (2001), *A Course in Time Series Analysis*, Wiley.

[19] Pennebaker, W. B., and Mitchell, J. L. (1993), *JPEG Still Image Data Compression Standard*, New York, NY: VanNostrand Reinhold, Chapter 4.

[20] Polikar, R. (1996), *The Wavelet Tutorial*,

`http://sun00.rowan.edu/ polikar/WAVELETS/WTtutorial.html` .

[21] Shenon, P. (2001) *U.S. Is Stepping Up Plan for Handling Anthrax Threat.* New York Times

[22] Shimizu K. (1995). Signal Detection Using the Wavelet Transform

[23] Stephenson J. (1997). Pentagon-funded research takes aim at agents of biological warfare. JAMA;278:(373-375).

[24] Vining, G. G. (1998), *Statistical Methods for Engineers*, Duxbury.

[25] Wagner M. et al (2000). Interim Report. The Nation's Current Capacity for the Early Detection of Public Health Threats including Bioterrorism. A review copy

[26] Wagner et al. (2001) The Emerging Science of Very Early Detection of Disease Outbreaks.

[27] Weigend, A. and Gershenfeld, N. (1993). Time Series Prediction: Forecasting the Future and Understanding the Past, Addison Wesley.

[28] Yu, P., Goldenberg, A. & Bi, Z. (2001) "Time Series Forecasting Using Wavelets with Predictor-Corrector Boundary Treatment", To appear in the *Proceedings of the Temporal Data Mining* Workshop at the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.