Using grocery sales data for the detection of bio-terrorist attacks

Anna Goldenberg[1], Galit Shmueli[1,2,*] and Rich Caruana[1]

[1] *Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213*

[2] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213*

SUMMARY

In this paper we explore the potential of using sales of grocery items data for early detection of epidemiological outbreaks and bio-terrorism attacks. These data are of special importance, as they are illustrative of non-symptom specific data that are expected to arrive earlier than medical data commonly used for such purposes. We explore the characteristics of such data and create a detection algorithm that detects irregular patterns of purchases that indicate an epidemiological outbreak. We show that it is feasible to use non-specific syndrome data, such as over-the-counter medication sales, for early detection of bio-terrorism attacks. Our conclusions are based on experiments with a theoretical simulation of a large anthrax outbreak. The proposed detection system consists of several layers and combines methods from signal processing, machine learning, statistics, and quality control. Finally,

*Correspondence to: Department of Statistics, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA. Phone: (412) 268-1880, FAX: (412) 268-7828, E-mail: galit.shmueli@cmu.edu

Contract/grant sponsor: The Centers for Disease Control and Prevention. The contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.; contract/grant number: U90/CCU318753-01

Contract/grant sponsor: The Agency for Healthcare Research and Quality; contract/grant number: 290-00-0009

we propose an evaluation scheme for testing a detection system when the data lack in outbreaks. To illustrate the framework and its evaluation, we apply the detection system to sales of over-the-counter cough and cold medication.

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 a population. The effect of introducing a bio-agent can be enormous, bringing deaths to many people and thus requires timely identification and proper response. It is desirable to detect an epidemic or an attack as soon as it occurs, but what are possible sources that react early in the presence of a public health threat, indicating the first signals of sickness in a large group of people?

Existing detection systems usually use as input data that are collected from medical sources (e.g., ER records and lab tests) or public health sources (e.g., school absence records). Although these types of data are direct measures of illness, in most cases they do not arrive in a timely manner. Upon exhibiting symptoms of illness, most people do not seek emergency medical services immediately but start by taking over-the-counter medications, by looking for symptoms on the web, by calling friends, etc. By the time the medical community has received a sufficient number of calls to get concerned about the spread of illness due to a bioterrorism attack, it might be too late.

One valuable source of timely data is over-the-counter (OTC) medication purchases information. This type of data tends to be very rich and large, and although it does not measure illnesses directly, it can infer specific symptoms that are being experienced by purchasers at a

relatively early stage of an outbreak. Although such data are typically noisy, the potential of discovering the first signals of an outbreak (or a bioterrorism attack) is promising. [1] showed that the potential of such data for the detection of a seasonal flu, that in turn might be an indication of a bioterrorism attack, can be enormous.

Figure 1 is a plot of several data sources over time. It includes three types of data: medical data including ER admissions divided into different diagnoses, mortality reports, and lab cultures; OTC medication sales; and other non-symptom specific data such as school absence records and web queries about various symptoms (cough, headache, etc). The sales are given on the graph on a weekly scale since some of the data were not available in daily form. Yet the graph reveals the potential relation between the different sources of data. It is clear that all the series peak around the same time.

In most case grocery are gathered typically for management and subsequently analyzed for marketing rather than epidemic detection purposes. The research, often sponsored by private corporations owning a particular supermarket chain, has concentrated on development of pricing strategies, analyzing pricing sensitivity (e.g., [3]), etc.

There is at least one project that used OTC sales for detection of an outbreak: The Diarrheal Disease Surveillance [4] which observed pharmaceutical data from 38 drugstores around New York area. The purpose of the work was to find out whether information collected about the sales of OTC medications could be a valuable source for rapid detection of outbreaks of diarrheal illnesses. The project confirmed the usefulness of pharmaceutical type data and suggested its use in the surveillance of other diseases. However, it proposed no clear framework.

In this paper we introduce an extensive and general framework for using grocery data for the detection of epidemic outbreaks or bioterrorism attacks. The detection system uses grocery

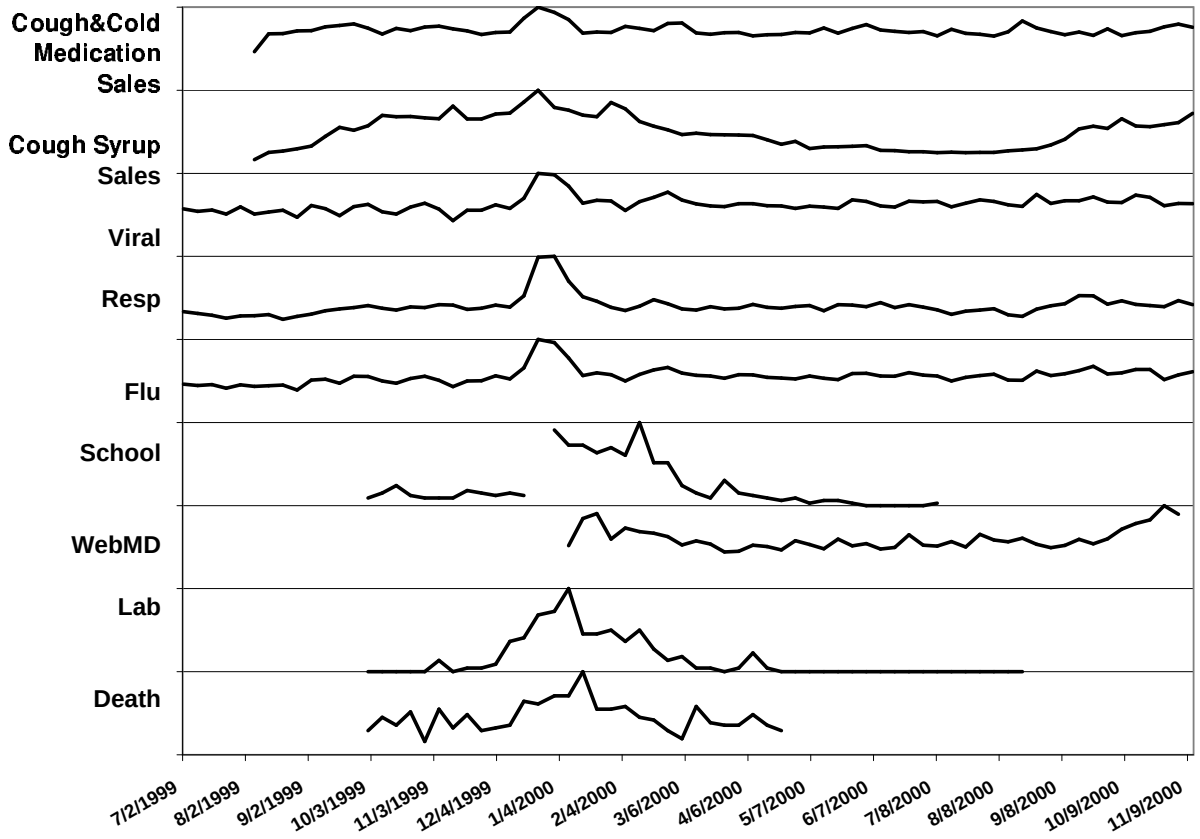

Figure 1. Comparing various data sources over time. From top to bottom: OTC medication sales, medical data, and other non symptom-specific data (reproduced from [2], with the addition of OTC medication sales).

sales data as input, and when it encounters an abnormality in sales that might indicate an outbreak, it signals a warning. The system is tailored to handle the special features of grocery-type data. Section 2 describes the main features of this type of data, and Section 3 introduces pre-processing methods that are suited to these special features.

The method developed in this project is based on the conditions and constraints of the problem at hand. Very little is known about the ways epidemics manifest themselves in grocery data. Even less is known about epidemics resulting from bioterrorism attacks. Hence, the system could not be constructed using an ordinary approach. Our suggested system gains its sophistication from using a combination of methods from various disciplines. Each of the methods, as well as their combination, is traceable in contrast with "black-box" algorithms such as neural networks. The output from a traceable mechanism is easier to interpret and relate to other information. Each of the system's layers is described in detail in section 4.

Following the detailed description of the detection system, we assess its performance. Because of the lack of known outbreaks and their manifestation in grocery data, the ordinary measures of sensitivity, timeliness, and specificity can not be used in their simple form. Instead, in Section 5, we suggest a modified methodology for testing the performance of a detection algorithm for a given dataset.

Throughout the paper we illustrate the use of the framework, applying it step by step to OTC cough medication sales. A discussion of the strengths and weaknesses of the detection system and some proposed future enhancements conclude this paper.

2. Properties of grocery data

Although most marketing research dealing with grocery data uses data that are aggregated on a weekly, monthly, or annual basis, this level of aggregation is not suitable as input for an early detection system. While purchases are usually recorded on a much finer time scale (e.g., minute accuracy) an over refined scale might result in data that are too noisy. A good balance is achieved by using daily-based data. Another reason for using daily data is for comparison with medical and public health datasets that are usually not available at a more refined level than daily records. We illustrate our framework using daily data, but it is general in the sense that it can be adapted for use with data at a more refined or less refined time scale.

Another issue concerning the refinement of the data is its form. Grocery datasets usually include data at the basket level. This means that each transaction is recorded with its entire details (exact date and time, customer information, detailed information on each product (SKU), prices, etc.). Although this information is available in the dataset, it might be inaccessible for use outside that chain. A grocery chain might be unwilling to disclose such information for many reasons. Also, such rich datasets are usually very large and require huge storage space, extra processing time, and excessive computing power.

A more compact form of grocery data is the aggregated daily sales of certain products over all baskets. This type of data is easier to obtain and to handle. The main information that is lost is the combinations of products that appear within a basket, which can be useful for detection purposes. Also, the number of items tracked need not be the total available items in the store. Optimally, the selection of items to be tracked is done by a group of experts from the fields of epidemiology, public health, and marketing.

We focus on daily aggregated data as input into our detection system, and show that even at

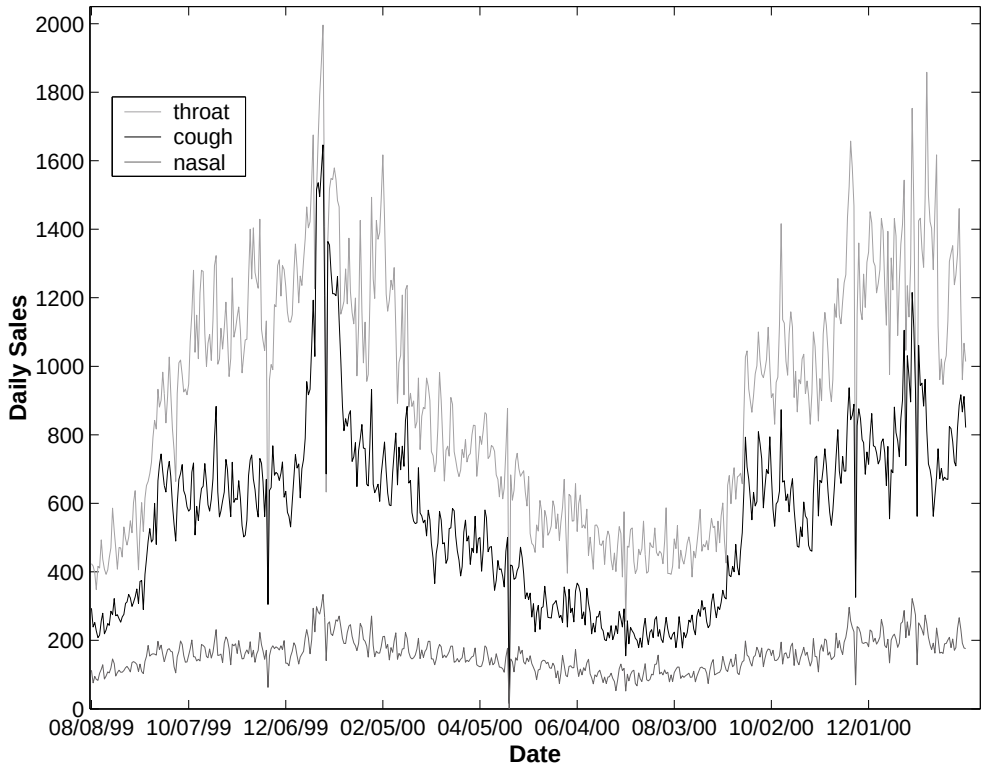

Figure 2. Sales for three OTC subgroups from 8/1999-1/2001

this level there is enough information for detecting abnormalities. The data that were available to us and used in the project are the daily sales of retail over-the-counter medication and a few grocery items that are known to be related to specific symptoms (e.g., facial tissues, orange juice, and soup). The data are aggregated for individual products over the period of a single day. The data span 541 days beginning August 8, 1999 until January 31, 2001. They consist of 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. The daily sales for three of the subgroups are described in figure 2.

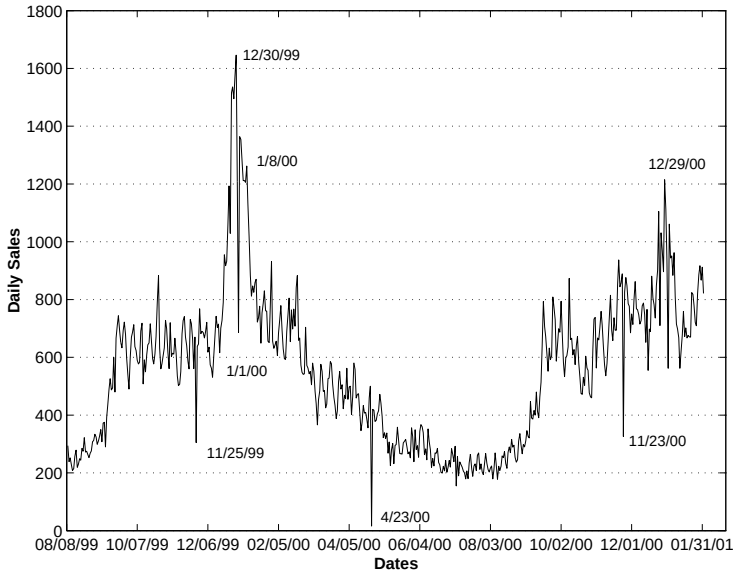

Figure 3. Sales of Cough Syrup/Liquid Decongestant from 8/1999-1/2001

In this paper we will use only the first subgroup of Cough Syrup/Liquid Decongestant to illustrate the proposed framework. For illustration purposes we use only the first 365 daily sales, and use the remaining 176 points for system evaluation (section 5). A plot of the sales for this subgroup over time, for the entire period, can be seen in Figure 3.

One of the dominant features of aggregated sales over time is its seasonal component. The sales of a certain product or category of products might vary considerably

- During different days of the week,
- From season to season,
- During holidays,
- When store opening hours change (e.g., on holidays).

In addition, there may be steady trends over time (such as a linear increase in sales), and temporary increases or decreases in sales for reasons other than an outbreak (e.g., new store policies, pricing).

Some of these effects can be seen for the sales of cough medication in figure 3. It is evident that there is a seasonal summer/winter effect and that the sales in winter are generally higher and more chaotic. It can also be seen that there are high peaks around winter holidays such as a huge spike between Christmas and New Year. There is also a drop almost to zero around April 24, 2000, which indicates that most stores were closed on Easter. Other visible effects are a 7-day periodicity and a weekday/weekend effect where sales during weekdays are generally lower than on weekends with the highest sales being on Saturdays.

3. Preprocessing grocery data

3.1. Scaling the data

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. Since daily and weekly variations are due to store-wide sales patterns rather than the fluctuations of a disease in the population, scaling the data is roughly equivalent to averaging the sales effects. The new scaled data are the proportional sales, rather than the raw counts of purchases. This scaling can help to eliminate most of the effects described above.

An alternative normalizing constant is the total daily store-wide sales. This piece of information was not available in our case, and we used the daily sales of a certain category as an approximation. Another option which we did not pursue was to use the sales of a relatively

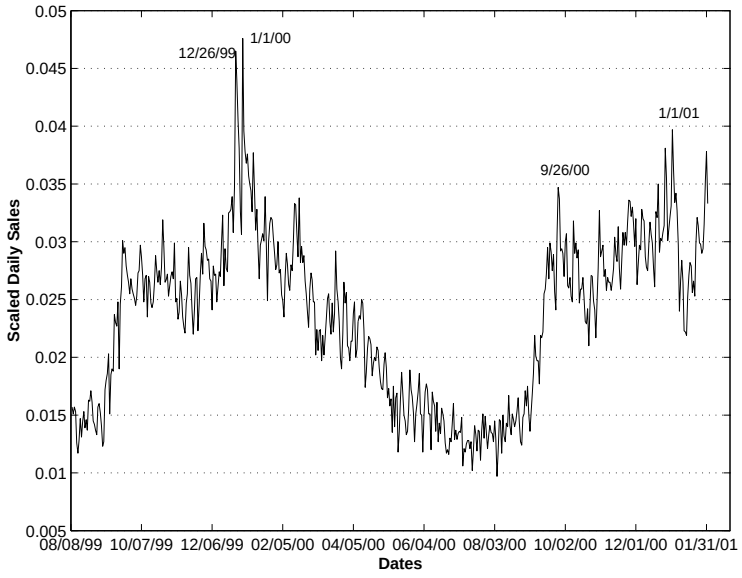

Figure 4. Sales of Cough Syrup/Liquid Decongestant, Scaled by Total Sales

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 these reasons we decided not to pursue this scaling option.

In order to scale the daily cough medication sales, we thus divide the daily sales by the total sales of all the products in our database, i.e., of all health related items. The new scaled series is plotted in figure 4. In comparison with the raw counts, the scaled data do not exhibit the holiday peaks, and they are generally smoother. From here on, we only consider these ratios.

3.2. Accounting for closed store dates

In figure 4 it can be seen that on three occasions the sales were nearly zero. These correspond to dates when most of the shops were closed. In order to eliminated the influence of this irregularity, these points were replaced by interpolations using the four points before and four points after the interpolated value. This means that a signal on such days should usually be ignored, or at least treated with caution.

3.3. The "Periodicity Plot"

Periodicities that are less obvious but suspected may also be present in the data. To detect such periodicities, we introduce a new graphic tool that removes seasonality components iteratively leading to a "periodicity plot". The idea is to remove from the data one seasonal component at a time, where the first component is a single day ($i = 1$), then every other day ($i = 2$), 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 the 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 implies visually which substantial periodic components exist in the data. These components can then be interpreted, and if sensible, removed. Figure 5 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.

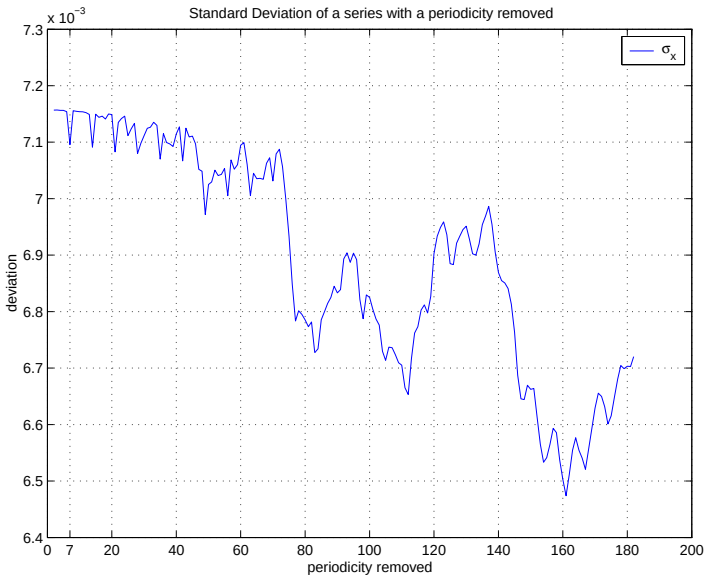

Figure 5. Periodicity plot for the scaled cough medication sales

4. The detection algorithm

The general structure of our framework consists of four steps: At first, we de-noise the data. Next, the de-noised data is used for predicting the next day's sales. Based on the predictions, we create an upper control limit. When the next day's sales data arrives, it is compared to the upper control limit, and if it exceeds it, the system flags, signaling that the new point is larger than expected. Next, we explain each of the four steps in detail.

4.1. De-noising the data using the Discrete Cosine Transform

The first layer of our proposed detection system is to learn the main features of the data. This knowledge is then used in the next layer to predict future sales. To increase prediction accuracy it is important that effects that are rare and chaotic in the data are eliminated. In

frequency domain, noise in the data is represented by weak and infrequent frequencies. The field of signal processing includes various noise reduction techniques, among which we chose to use the Discrete Cosine Transformation (DCT) to de-noise our data. Unlike most of the other techniques that are aimed at modeling the localized signal, 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.

DCT is very closely related to Discrete Fourier Transform (DFT). In fact, the only major difference is that the cosine transform approximates the function using cosines alone rather than complex cosine-sine functions. This allows us to stay with the real scale. To reconstruct the original series, the inverse discrete cosine transform (IDCT) is used (for more details, see appendix). The DCT technique has the advantage that a sequence can often be reconstructed very accurately from only a few DCT coefficients. This is a useful property for applications requiring data reduction as in the case at hand. We use the term "de-noised data" to denote the reconstructed series from using IDCT.

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 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 the usual method used - high-pass/low-pass, or "vertical" filtering, which 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 model as close to the data as possible, for the purpose of accurate forecasting. Figure 6 contrasts the effect of removing coefficients below

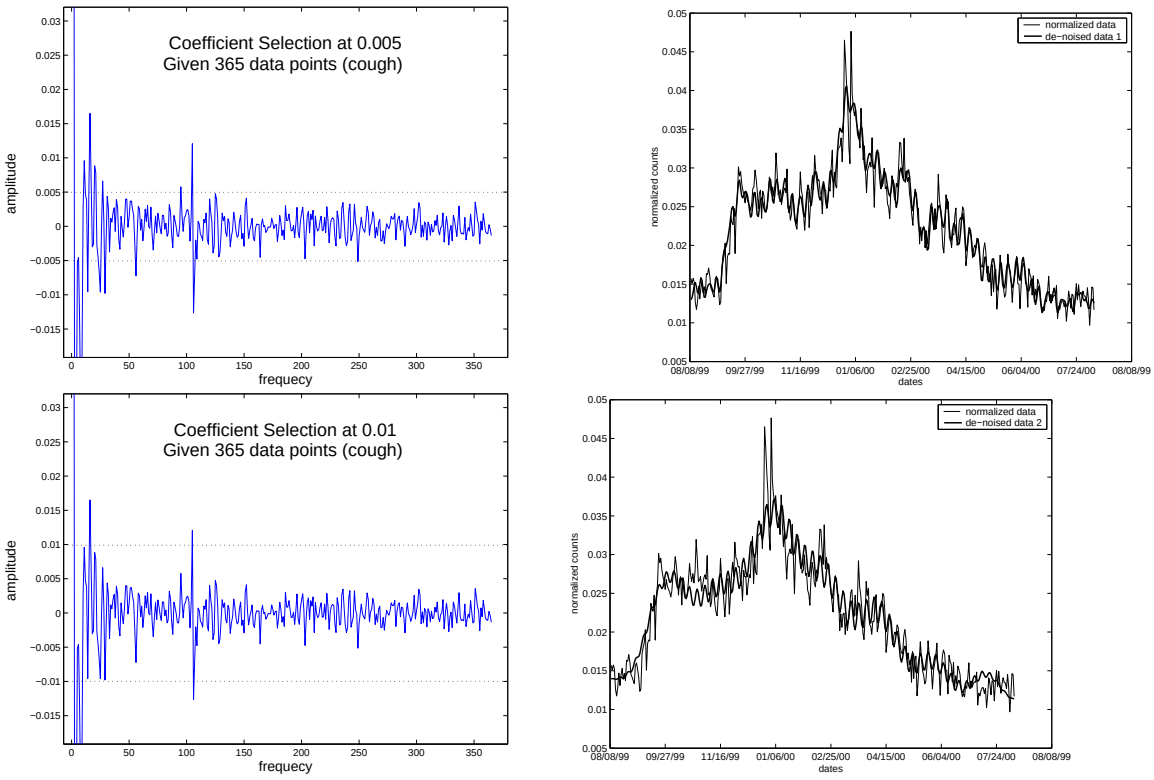

Figure 6. The effect of eliminating DCT coefficients below a certain threshold (on the left) on the reconstructed data (on the right), for a threshold of 0.005 (top graphs) and of 0.01 (bottom graphs)

a 0.005 threshold compared with a 0.01 threshold, along with the series that are reconstructed from the reduced set of coefficients. Figure 6 illustrates the general result: that the lower the threshold, the higher the correspondence between the original data and the reconstructed series.

To evaluate the effect of removing coefficients on the closeness of the de-noised data (denoted

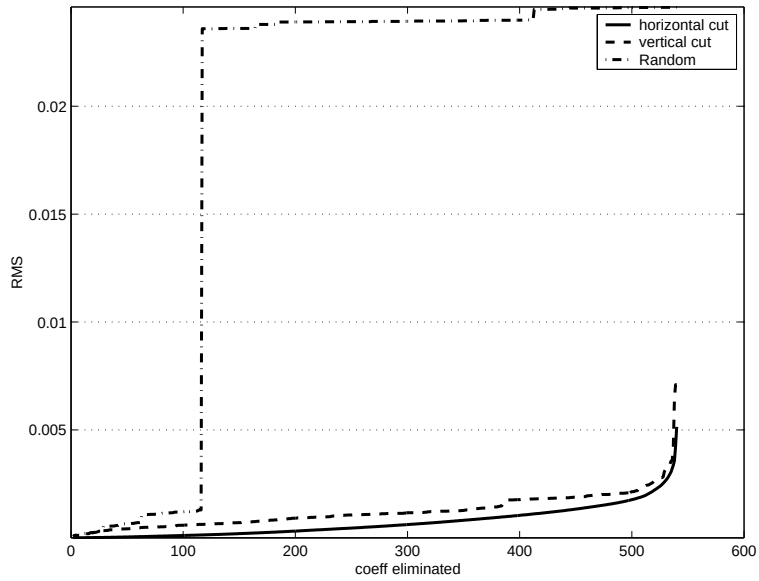

Figure 7. The effect of eliminating DCT coefficients on the RMS(de-noised, scaled) measure for three types of elimination: using a horizontal, vertical, and random threshold

by D_i) to the scaled data (denoted by X_i), we define a Root Mean Square (RMS) measure for N data points, as:

$$RMS(\text{de-noised, scaled}) = \sqrt{\sum_{i=1}^{N}(D_i - X_i)^2} \tag{1}$$

Figure 7 illustrates the effect of coefficient removal on the RMS measure. The closer the de-noised or reconstructed series is to the original one, the smaller is the RMS.

The third method displayed in figure 7 is random coefficient elimination; it was added to show that it does indeed matter which coefficients are removed. It can be seen that coefficient removal using horizontal filtering (represented by the solid line) has the smallest RMS of the three methods, for any given number of coefficients used to reconstruct the data. In addition, it can be seen that the horizontal cut is more robust in the sense that it has a less drastic effect on the RMS. This means that the addition or removal of a coefficient still results in smaller

error relative to the other methods.

The threshold for removing DCT coefficients is selected according to predictability and goodness of fit criteria, rather than setting it to a certain level. Each threshold selection leads to a tradeoff between simplicity and goodness of fit. The more coefficients one removes, the simpler and more predictable the de-noised series becomes. In fact, if one removes all but the first coefficient, the de-noised series is represented by a straight line. The de-noised series is then easy to predict but the farthest from the data in the RMS-error sense. To select the level of de-noising that approximates the data reasonably well and still is flexible for predicting new observations, we use cross-validation, which is a statistical tool widely used in machine learning: We separate the existing N data points into a training set of n_1 data points and a prediction set of n_2 data points. The method proceeds as follows: At first, we de-noise the n_1 observations or training data, and measure the difference between the data and the de-noised data to assess their closeness, by taking the RMS of the differences. We denote scaled observations by X_i and de-noised observations by D_i. The measure of closeness is then given by

$$RMS_{training}(scaled, de\text{-}noised) = \sqrt{\sum_{i=1}^{n_1}(X_i - D_i)^2} \tag{2}$$

As more DCT coefficients are removed, we expect this RMS to increase. Next, using a one-step ahead linear prediction (AR(1)) point n_1+1 is predicted from the previous de-noised data, and denoted by $\hat{D}_{n_1+1}$. We then continue to predict points $n_1+2, n_1+3, \ldots$ using a roll forward one-step prediction, to obtain $\hat{D}_{n_1+2}, \hat{D}_{n_1+3}, \ldots$. The discrepancy between these predictions and their corresponding real values in the prediction set is measured by

$$RMS_{prediction}(scaled, prediction) = \sqrt{\sum_{i=n_1+1}^{N}(X_i - \hat{D}_i)^2} \tag{3}$$

This RMS is expected to increase as more DCT coefficients are removed. At the third step,

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 $n_1 + 1$ 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

$$RMS_{prediction}(\textit{de-noised}, \textit{prediction}) = \sqrt{\sum_{i=n_1+1}^{N} (D_i - \hat{D}_i)^2} \tag{4}$$

As opposed to 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 a model that best fits the data under the constraint of having high prediction accuracy. In figure 8 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 these 21 coefficients.

4.2. Forecasting next day sales, using autoregressive models and Wavelet decomposition

Forecasting models are generally 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 [5]. They are simple to understand and interpret, straightforward to implement, and have been predominant forecasting tools for more than 50 years [6]. Autoregressive (AR) and Moving Average (MA) models are included in this category, as well as the more general ARMA models. The main disadvantage of linear models is that they are inappropriate for modeling even moderately complicated series, and have encountered

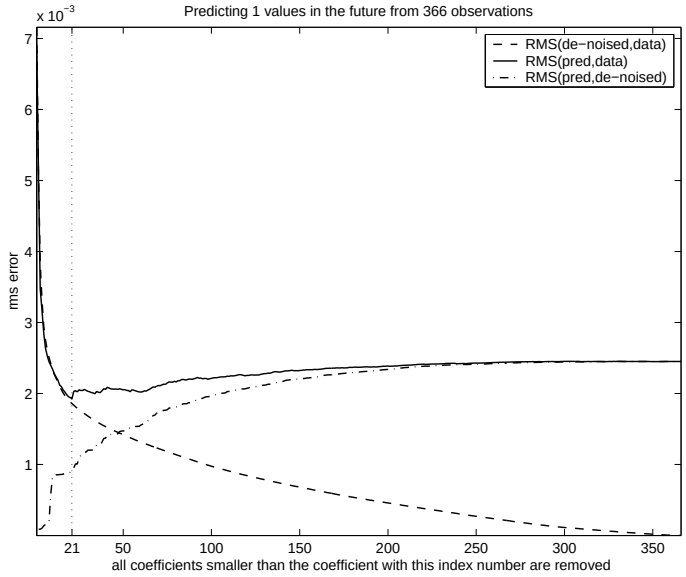

Figure 8. Minimizing the three types of RMS errors for the cough medication sales. The sum of the three is minimized at 21 coefficients.

various limitations in real applications [7]. 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 [5]).

From initial experimentation with grocery data, we learnt that linear models such as ARIMA, when applied to the scaled data, are not able to capture the complexity of the data's structure, and they 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 [8]). 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 or non-linear model to the scaled or de-noised data directly: since the preprocessing 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 [9] and is based on wavelet decomposition and prediction 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 separately and the individual predictions are recombined to form the final forecast. Unlike [9] who used neural networks for prediction, we use simple linear models [11].

The wavelet transform is a synthesis of ideas from engineering, mathematics, and physics, and is well integrated into a variety of software packages using an efficient algorithm [12]. The Discrete Wavelet Transform (DWT) is similar to DFT or DCT. 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 [13]. Second, it is suitable for use with non-stationary data [12]. Third, it often requires less coefficients to capture the main features of the series than in classical Fourier analysis. These advantages are especially meaningful when designing an automated system.

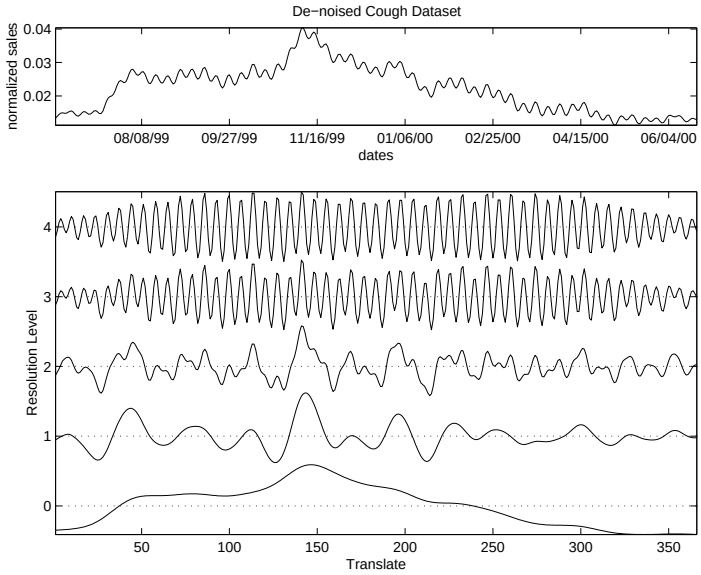

Figure 9. Decomposing the data into five resolutions using the redundant wavelet transform

To predict the next day sales, we use the de-noised series. This series is then 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 more details, see appendix). Figure 9 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 model. We use the AR model, which belongs to the category of linear models, for reasons of simplicity and interpretability. 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

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \epsilon_t \quad t = 1, 2, \ldots \tag{5}$$

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 [14].

In comparison to the complicated structure of the de-noised series, the wavelet coefficients within a certain resolution create a 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 the de-noised sales. Figure 10 illustrates the prediction for each of the resolutions separately, and their sum which is the prediction of the next day (de-noised) sales.

4.3. Setting an upper control limit on predictions

The algorithm for detecting an abnormal sales volume is based on making a one-step prediction using the de-noised data, and then comparing it to the actual value of the next day (de-noised) sales. If the actual sales are much higher than the predicted sales, the algorithm signals an alarm. The problem is to decide on the size of the discrepancy that represents a real abnormality, rather than just ordinary variation. In the quality control literature, a well-known method for monitoring a process for abnormalities is to use control charts. In its simplest form, the control chart has an upper and/or lower control limit, which if exceeded by a new data

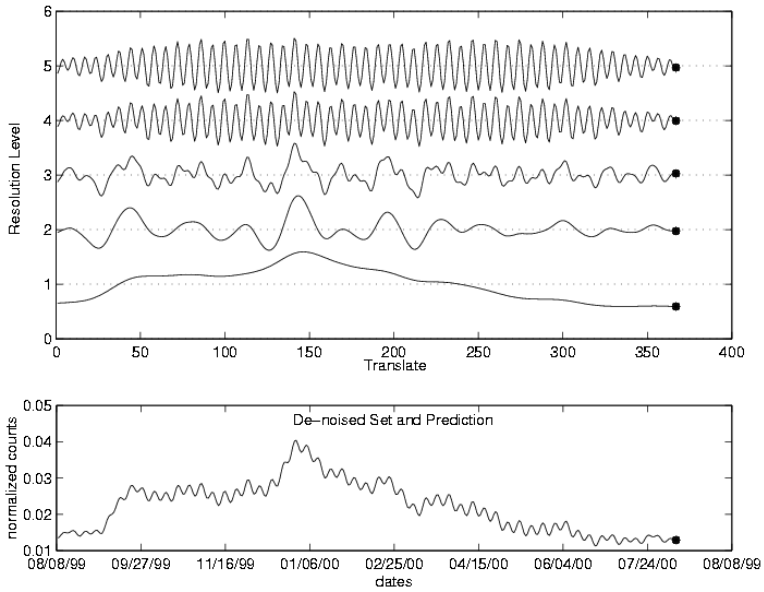

Figure 10. One step ahead predictions for each resolution and their sum

point, signals an alarm [15]. Using this approach, we can translate the discrepancy between the predicted and real sales volume using control chart formulation, by setting an upper control limit on the predicted value. If the actual value exceeds this control limit, then our system signals an alarm.

Specifying the upper control limit (UCL) is one of the critical decisions that must be made in designing a framework. For example, by making the UCL closer to the prediction we increase the risk of a type I error, or false alarm, which is the risk of a "regular", non-outlier point signaling an alarm. On the other hand, this will decrease the type II error, which is the risk of not identifying a true outlier. In the context of public health, the type I and type II errors respectively 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 risking the lives of many people. Since public health and government money are at stake, the

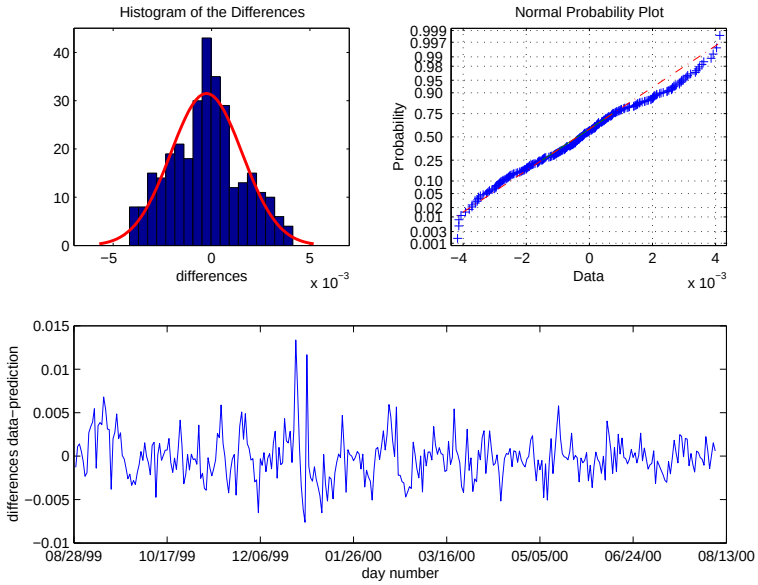

Figure 11. Checking the Normality Assumption of the differences between predictions and real data

tradeoff of the losses needs to reflect some consensus regarding the public's utility, and is not in the domain of the system designer.

For a given set of type I and II error probabilities we consider the distribution of the differences between the original data and their predictions for setting the UCL. We assume that the distribution of the differences is approximately normal, since the differences are a sum of errors that result from the de-noising and forecasting steps, and should therefore approximately follow a normal distribution. To check this assumption for specific data we use statistical methods such as histograms, probability plots, and more formal statistical tests. Figure 11 illustrates the use of graphical methods for assessing the normality assumption for the cough medication sales.

From experimentation with real data we learned that there might exist a very small percentage of differences that are extremely large (more than one would expect in a normal

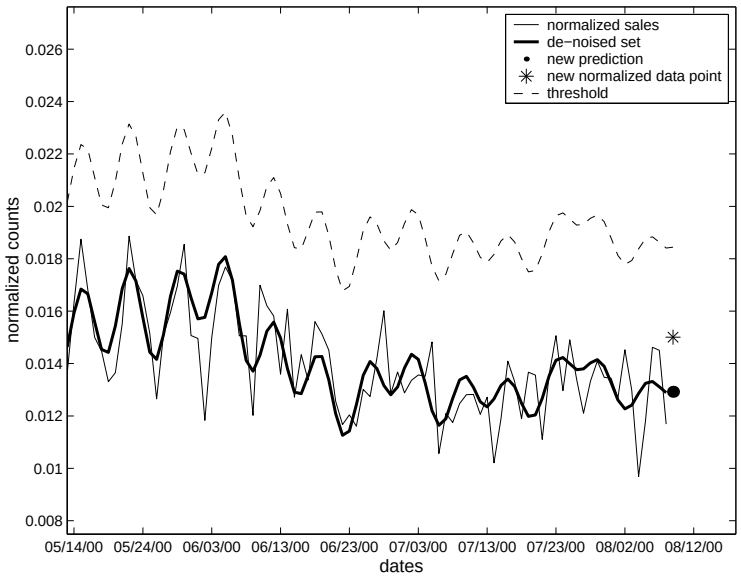

Figure 12. Comparing the next day prediction to the upper control limit for cough medication data

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 UCL selection process is then very similar to the method used in control chart literature [15]. In control charts it is customary to select the UCL to be some multiple of the standard deviation of the distribution. It is also common practice to select 3 as the coefficient. In other words, the UCL is the expected value plus 3 standard deviations. A 3-sigma upper limit will give a false alarm (type I error) probability of 0.001 if the normality assumption is valid. This type of chart is known to perform well in practice.

Figure 12 illustrates the process of comparing a new data point to the 3-sigma upper control limit for the cough medication data. The new data point does not exceed the control limit, as shown on the figure, and is therefore considered an ordinary point.

5. Evaluating the algorithm

5.1. Ordinary methods

A good early detection system is one that detects real abnormalities quickly, and does not signal any false alarms. These two features of sensitivity to real signals and insensitivity to false ones can not be achieved fully in practice, but can serve as guidelines for evaluating a system and for comparing systems. In order to evaluate how sensitive a system is to a real signal, the test data must include such signals. If we know

1. where in the data the signa/s occur and
2. the manifestation of the signal in the data (the structure of the signal),

then we can evaluate the system's ability, or power, to detect real signals by measuring whether the signal was detected, and if so, the time (or number of data points) until detection. In the medical and epidemiological literature this is related to the notion of *sensitivity*, i.e. the rate at which real alarms are correctly detected.

To determine the system's false alarm rate, we must know which data points do *not* include a real signal. A measure that reflects the tendency of a system to signal false alarms is the number/percent of alarms that were flagged at points that did not include a real signal. This relates to the term *specificity* used in epidemiology and medicine, which is the number of false alarms not flagged by the system divided by the total number of data points that do not include a real signal.

The quality control literature uses a different measure, which takes into account the timeliness of the detection. The *run length* is a random variable that counts the time until a flag is raised [16]. If the flag was raised during a real event then the run length measures the

time until detection. If the flag was raised on a non-event then the run length measures the time until a false alarm is encountered. A combination of sensitivity, specificity, and run-length can give the joint information of "if and when" a real signal is detected, or "if and when" a false alarm occurs. In addition, when the points are independent the information contained in the measures of specificity (or sensitivity) and run-length is equivalent. However, when there exists dependence in the data, the run-length is more useful in the sense of interpretability.

When the signal is a single spike in the data, the computation of sensitivity and specificity is straightforward. However, a complication arises when the signal spans over a longer period than a single point in time. For example, a certain symptom can manifest itself as a rise in sales of a certain product over a period of three days, and then return to its ordinary sales volume. If a detection system flags on the fourth day, the run-length approach would consider this event a lagged detection. However, for the computation of the sensitivity and specificity, is the flag on the 4th day considered a false alarm or a detection with a lag of 4 days? It is thus necessary to define in advance what a false alarm is as opposed to a lagged detection.

5.2. Assessing performance without known outbreaks

In many real world cases, and especially when considering bio-terrorism attacks, there is no information about the manifestation of a signal in the data. In the case of anthrax, for example, there is no available data representing the course of anthrax on the sales of OTC medication or grocery items. This means that the ordinary way of evaluating the power of a system to detect an anthrax outbreak can not be used. The only piece of information that can be assumed is that the data do not contain a signal that indicated an anthrax outbreak. Thus, the false alarm rate can be measured, but not the detection of real alarms, which is the goal of this project. For

epidemiological outbreaks the situation can be even worse. For example, to assess the ability of a system to detect a flu outbreak there must be a clear identification of flu periods. It turns out that information on exact dates of flu outbreaks in a restricted geographical region (e.g. the Allegheney county) is not readily available. In fact, there is very little reliable data about any diseases that occurred in the Pittsburgh area in the past year and a half. This means that the ability to evaluate both the false alarm and real alarm detection can not be done using the ordinary method.

[18] 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.

We propose an alternative method for evaluating a detection algorithm, which is more general and uses real data rather than simulated data. This allows for dependent data (which is often the case), and outbreaks that span over several days as opposed to occuring at single point in time. We take a supervised learning to approach to evaluation. The idea is to add a pattern that is a combination of spikes, and represents the footprint of a specific outbreak, directly into the data. The construction of this footprint is done in collaboration with a group of experts (from the fields of epidemiology, medicine, and marketing). Figure 13 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 are inhaled, it causes fever, difficulty of breathing, and death can occur within days. Based on this information, and with the assistance of an

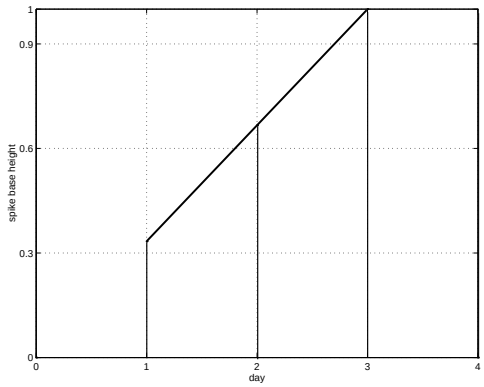

Figure 13. Hypothesized structure of the footprint of an anthrax outbreak in the sales data

expert of medical informatics from the Center for Biomedical Informatics at the University of Pittsburgh, we designed a pattern that spans three days with a linearly increasing rate of sales. For a detailed discussion of OTC medication sales in the context of detecting Anthrax, see [17].

After the structure of the signal is specified, it is added sequentially at *every* point of the data (unlike the method by [18] where the spike is added to one specific point in time). For reasons mentioned above, we treat all data points that are "far enough" from data that contain the added signal as "clean data" with no signal. The addition of a simulated anthrax footprint in the cough medication data is illustrated in figure 14. 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.

The system can now be tested on the modified data and its detection power can be assessed. However, the ordinary measures of sensitivity, specificity, and timeliness carry a different meaning. We propose a measure that estimates the probability of detecting a specific pattern:

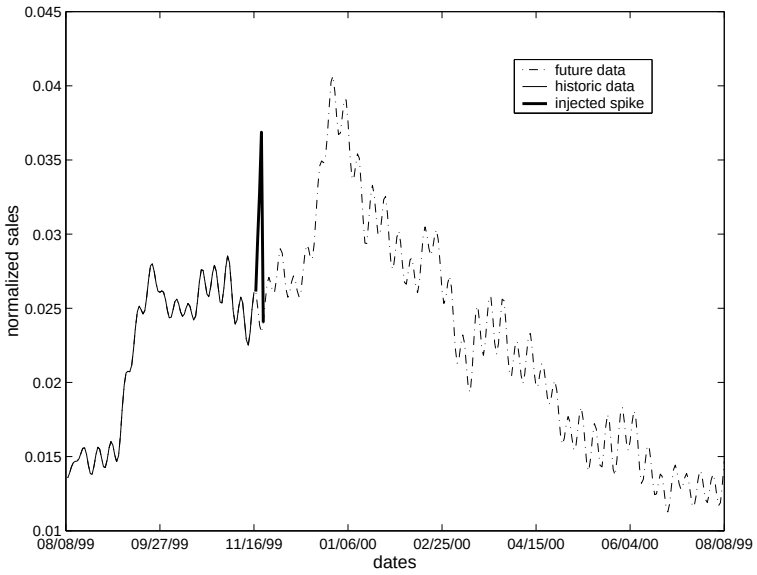

Figure 14. Adding the footprint into the daily sales data: the data with a 3-day pattern added on 11/16/99

the Spike Detection Ratio (SDR):

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

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

This is the proportion of spikes detected divided by the total spikes added. 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). This is under the assumption that the pattern is added sequentially at every point in time, 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. Figure 15 illustrates the SDR within a single day, two days, and three days when a shape similar to that in figure 13 and slope of $1/3$ (i.e. a pattern of heights $[1/3, 2/3, 1] \times constant$) is

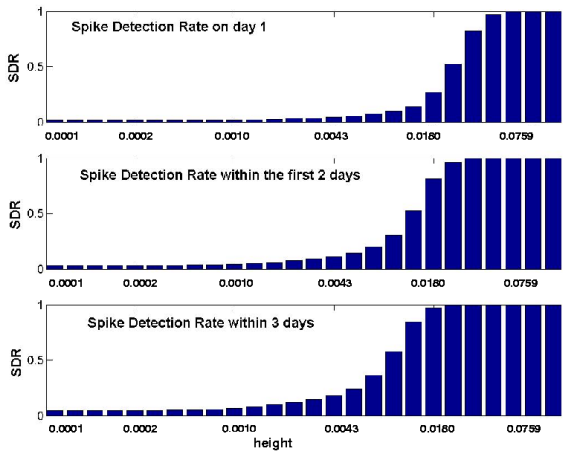

Figure 15. Spike Detection Ratio as a function of the pattern's height, within a single day, two days, and three days. The slope of the pattern is 1/3.

added to the data. The SDR is a function of the pattern's height (the horizontal axis), which represents the amount added to the data. As the height increases, the pattern is detected more easily, leading to an increased SDR. The pattern was added to every data point, beginning on the 33rd point (the first 32 points were used to initialize the forecasting component).

Creating an evaluation scheme also enables us to tune the system and to try out different configurations. We illustrate this for the cough medication sales, where we compare the three configurations of the system that performed best from a set of 8 configurations, using the SDR evaluation scheme. Figure 16 compares the SDR of three systems:

1. DAR1 is the simplest configuration where DCT was not used to de-noise the data, and an AR(1) model was fit directly to the scaled data for predicting one-step ahead sales.
2. MAR7 is a system that applies DCT to the scaled data, and then uses the de-noised data to predict the next day sales, via an AR(7) model (without wavelet decomposition).
3. MWAR7 is the most sophisticated configuration, where the scaled data are de-noised,

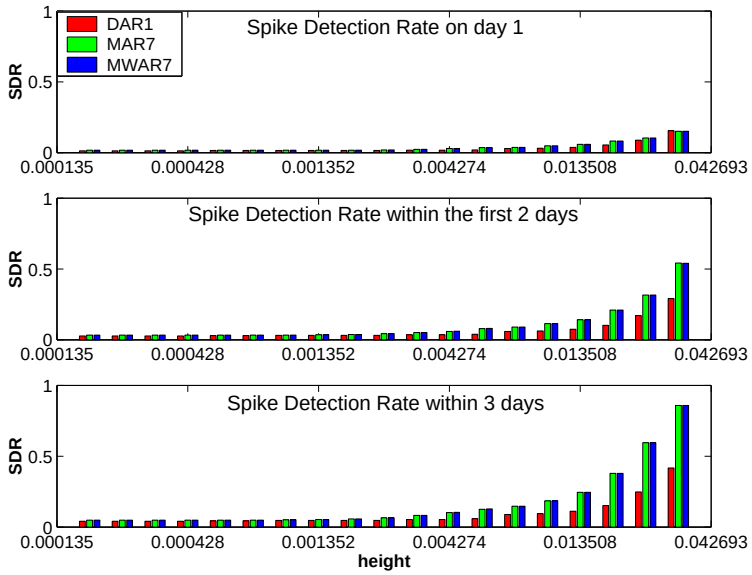

Figure 16. Comparing three configurations of the detection system using the SDR evaluation system

and then decomposed into wavelets. The wavelet components are used to forecast the next day sales using an AR(7) model.

For these data the last two models (MAR7 and MWAR7) perform equally well for a first day detection, as well as for detection within the first two or three days. Both configurations outperform the simple DAR1 model, which is too simplistic for these data.

In addition to testing different configurations of the system, by applying this evaluation scheme it is possible to learn about different aspects of the data and the framework. Firstly, it is possible to reveal characteristics of the data that are most sensitive to the insertion of a pattern (e.g. during weekends vs. weekdays, during periods of increases vs. decreases of sales). 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 length, 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.

6. Future detection system enhancements

The detection system that is described here uses a single data source as its input. Although sales of OTC medications tend to give an earlier warning of an attack than medical data, it would be highly useful to incorporate other data sources into our system. In order to integrate such data sources, their relation should be investigated. Based on the initial relation that is seen in Figure 1 we believe that an integration of the different data sources into the detection system is important for better detection as well as for better interpretability. A comparison on a daily scale would also reveal the possible lags between the manifestations of an epidemic footprint in the different series.

Although the proposed system is automated, there are two components that are tuned manually. These include the number of wavelet components selected and the order of the Autoregressive models. In a fully automated system, we would suggest to add criteria for automatically determining these parameters.

7. Evaluating the detection system

The proposed detection system described in this paper is a modular one. It is composed of several layers: First the data is preprocessed and rescaled. Next, it is de-noised. Then, a one-step ahead prediction is made and based on that an upper control limit is constructed. Finally, the system is evaluated using the SDR scheme. Any of the methods used in these layers (e.g., DCT, Wavelet decomposition, AR models, 3-sigma control limit) can be replaced

by other methods, or eliminated. This flexibility allows the detection system to be customized to the type of data, enabling the use of other non symptom-specific data as input into our detection system. From a preliminary investigation of other OTC medication sales we found that different configurations of the system were better for detection than the one used for the cough medication [10].

Our approach has been an empirical one. We selected the optimal configuration and tuning the parameters according to empirical measures. In addition, we used some well known tools in a novel way required by the problem at hand. For each new application of a known method, we explained the rational and evaluated its performance using empirical measures. We have not, however, investigated these uses theoretically. A theoretical evaluation of the system would in principle include:

1. Investigating the effects of applying DCT to data, and the coefficient selection procedure. Although the DCT does not attempt to model the data, the criteria for coefficient selection are goodness of fit and accuracy of prediction. The statistical properties of the proposed RMS-based method should be studied, and compared to other feasible goodness of fit measures.
2. Learning about the combination of wavelet decomposition and linear forecasting models (ARMA) from a theoretical viewpoint. ARMA models have been applied to wavelets in [11], in the same way that we have used them (predicting each wavelet component separately and combining the predictions to obtain the next point in the series). However, this combination has not been studied there theoretically.
3. Comparing the performance of the existing system to one that excludes the wavelet decomposition step. Although the empirical results are in favor of the wavelet

decomposition step even in combination with a simple AR model, its contribution to the accuracy of predictions and to the complexity of the system should be investigated and compared with a framework that models the de-noised data directly using ARIMA models.

4. The normality assumption underlying the control limit was evaluated using empirical tools. It is based on the rational that the differences should follow the Central Limit Theorem, since they are a sum of various errors. However this assumption remains unproven.

Appendix: Technical details

Discrete Cosine Transform

Mathematically, for an input sequence $x(n), n = 1, \dots N$, the result of the DCT transform is a vector of coefficients $y(k)$:

$$y(k) = w(k) \sum_{n=1}^{N} x(n) cos \frac{\pi(2n-1)(k-1)}{2N} \quad , k = 1, \dots, N \tag{7}$$

where

$$w(k) = \begin{cases} \sqrt{1/n} & , k = 1 \\ \sqrt{2/n} & , 2 \leq k \leq N \end{cases} \tag{8}$$

To reconstruct the original series from the DCT coefficients the Inverse-DCT (IDCT) is used, mathematically described as:

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

and $w(k)$ is defined as above. For further details on DCT and IDCT, refer to [19] and [20].

Discrete and Redundant Wavelet Transform

The basic 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 [21]. 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*):

$$\phi_{j,k}(t) = 2^{-j/2}\phi\left(\frac{t-2^j k}{2^j}\right) \quad (10)$$

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

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 [21]. 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 [22].

The 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) wavelet transform 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 [22].

In our application we have five resolutions: the first four represent the high frequencies in the data (detail coefficients) and the last, or the residual, reflects the low frequencies.

ACKNOWLEDGEMENTS

We thank Stephen Fienberg who was the principle investigator on our component of each of these projects, for his useful advice and valuable input. We also thank Dr. Michael Wagner for providing public health orientation and framing, and Dr. Jeremy Espino for assisting with collecting background information.

REFERENCES

1. Goldman L. "Inhalation Anthrax" in *Cecil Textbook of Medicine*. NY: Saunders Company, 2000.
2. Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, McGinnis LF, Deerfield DW, Druzdzel MJ, Fridsma DB. The Emerging Science of Very Early Detection of Disease Outbreaks, *Journal of Public Health Management Practice* 2001; **7**: 51-59.
3. Boatwright P, McCulloch R, Rossi P. Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model. *Journal of the American Statistical Association* 1999; **94**, 448:1063-1073.
4. Mikol YB, Miller JM, Ashendorff A. Diarrheal Disease Surveillance Programs: New York City's Experience, 2000. Presented at *The International Conference on Emerging Infectious Diseases* (also availble at `http://www.ci.nyc.ny.us/html/dep/pdf/wdrap00.pdf`, accessed Dec 12, 2001)
5. Chatfield C. *The Analysis of Time Series*. Chapman & Hall, 1989.
6. Weigend AS, Gershenfeld NA. *Time series prediction: Forecasting the future and understanding the past*. Addison-Wesley : Menlo Park, CA, 1993.
7. Peña D, Tiao GC, Tsay RS. *A Course in Time Series Analysis*. Wiley, 2001.
8. Brockwell PJ, Davis RA. *Introduction to Time Series and Forecasting*. Springer, 1996.
9. Aussem A, Murtagh F. Combining Neural Network Forecasts on Wavelet-Transformed Time Series, *Connection Science* 1997; **9**(1): 113-121.

10. Goldenberg A. Framework for Using Grocery Data for Early Detection of Bio-terrorism Attacks. Technical Report CMU-CALD-01-101, Carnegie Mellon University, 2001.
11. Yu P, Goldenberg A, Bi Z. Time Series Forecasting Using Wavelets with Predictor-Corrector Boundary Treatment. *Proceedings of the Temporal Data Mining Workshop at the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2001 (to appear).
12. Abramovich F, Bailey T, Sapatinas T. Wavelet analysis and its statistical applications. *The Statistician - Journal of the Royal Statistical Society, Ser. D* 2000; **49**: 1-29.
13. Polikar R. *The Wavelet Tutorial* 1996; `http://sun00.rowan.edu/ polikar/WAVELETS/WTtutorial.html` (last accessed Dec 12, 2001).
14. Hamilton J. *Time Series Analysis*. Princeton University Press: Princeton, New Jersey, 1994.
15. Montgomery DC. *Introduction to Statistical Quality Control*. Wiley, fourth ed., 2001.
16. Vining GG. *Statistical Methods for Engineers*. Duxbury, 1998; p. 267.
17. Goldenberg A, Shmueli G, Caruana R, Fienberg SF. Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales. *Proceeding of the National Academy of Sciences* 2002, to appear.
18. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreeaks of Infectious Disease. *Journal of the Royal Statistical Society A* 1996; **159**(3): 547-563.
19. Jain AK. *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
20. Pennebaker WB, Mitchell JL. *JPEG Still Image Data Compression Standard*, New York, NY: VanNostrand Reinhold, 1993; Chapter 4.
21. Shmuway RK, Stoffer DS. *Time Series Analysis and Its Applications*. Springer Verlag, 2000.
22. Nason GP, Silverman BW. The Stationary Wavelet Transform and Some Statistical Applications. In *Wavelets and Statistics, Lecture Notes in Statistics 103* Antoniadis A. & Oppenheim G. (eds), New-York: Springer-Verlag, 1995; pp. 281-300.

