A Wavelet-based Anomaly Detector for Early Detection of Disease Outbreaks

Thomas Lotze LOTZE@MATH.UMD.EDU
Galit Shmueli GSHMUELI@UMD.EDU
The Applied Math and Scientific Computation Program and The Robert H Smith School of Business, University of Maryland College Park, MD 20742 USA

Sean Murphy SEAN.MURPHY@JHUAPL.EDU
Howard Burkom HOWARD.BURKOM@JHUAPL.EDU
The Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723 USA

Abstract

We describe a wavelet-based automated algorithm for detecting disease outbreaks in temporal syndromic data. We describe the method, which improves upon the Goldenberg et al. (2002) algorithm and its implementation on a diverse set of real syndromic data from multiple data sources and multiple geographical locations. Our results show a robust performance which is comparable to a few recently suggested methods.

1. Introduction

We examine the efficacy of using wavelets to predict the behavior of authentic syndromic time series data and extract anomalies from normal operating patterns. Syndromic surveillance involves the monitoring of time series containing syndromic data, such as emergency room counts, pharmaceutical sales or doctor appointments, in order to detect disease outbreaks before traditional sentinel methods. Although syndromic surveillance has been developed using multiple techniques, there still remain issues of preconditioning the data to remove non-disease outbreak patterns, as well as finding alerting mechanisms triggered by a variety of outbreak patterns while maintaining an operationally acceptable false alarm rate.

One way of removing the non-outbreak-related patterns is to monitor the residuals obtained by subtracting predictions from observations. In theory, such residuals should yield random noise plus the outbreak signal, making the outbreak potentially more detectable by standard control chart methods (for more details on control chart methods, see Ryan (1989)). However, other techniques to detect outbreaks can also be applied directly to the non-normalized data (although most such techniques normalize internally).

Wavelets provide a promising means for developing both of these types of technique because they can decompose a signal at multiple time and frequency scales. This capability makes them a good candidate to handle syndromic data, which are created by processes operating on different time scales (from a weekend sale at a pharmacy to a 6 month flu season to a short term bioterrorism event) and by periodic effects at different frequencies (weekly, monthly, or yearly). In addition, wavelets are computationally tractable and can be easily modified for use in real-world applications.

Recent papers by Bakshi (1998) and Aradhye et al. (2003) have investigated the use of wavelets in statistical process charts for chemical engineering processes, and these investigations have inspired their application to the statistical process charts commonly used in syndromic surveillance. But while wavelets have been suggested by other researchers in syndromic surveillance (Shmueli and Fienberg, 2006), they have rarely been directly compared to other methods on real syndromic data. Goldenberg, et al. (2002) performed an analysis using wavelet predictions as a way of detecting a simulated anthrax outbreak. Wavelets are also used to some extent in the commercial RODS application, which uses the wavelets' averaged level to normalize for long-term trends and negative singularities (Zhang et al., 2003). Stacey et al. (2005) report a retrospective study of over-the-counter medication sales using wavelets to better understand trends and patterns. In line with the Goldenberg et al. (2002) implementation and in contrast to Zhang et al. (2003), we introduce two preconditioning steps to deal with day-of-week effect and holidays, and then use all levels of the wavelets in order to either predict or alarm.

In this paper, we empirically examine the performance of this algorithm and several modifications, and compare it with other methods for removing non-outbreak related signals.

Appearing in *Proceedings of the 23nd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

2. Data

This study used time series of aggregated, de-identified counts of health indicators derived from the BioALIRT program conducted by the U.S. Defense Advanced Research Projects Agency (DARPA) (Siegrist and Pavlin, 2004). Three types of daily syndromic counts were represented: military clinic visit diagnoses, filled military prescriptions, and civilian physician office visits. The BioALIRT program categorized the records from each data type as Respiratory (RESP), Gastrointestinal (GI), or Other, and the data were gathered from ten U.S. metropolitan areas with substantial representation of each data type. This study used the RESP and GI data for all three data types from five of the cities for a total of 30 time series, each including syndromic counts for 700 days. The RESP time series showed consistent day-of-week effects and seasonal trends, while the GI time series showed only day-of-week effects. Figure 1 shows two sample series, one of RESP (top) and one of GI (bottom) data from one particular city.

To restrict attention to authentic data representing routine consumer behavior and disease trends, but not containing artifacts such as changing participation of data providers, we excluded time series in which temporary dropouts and permanent step increases were evident. While an operational health-monitoring system must manage such data problems, the goal of this investigation was to isolate the issue of removing systematic data behavior from these problems and from the choice of alerting methods that use the data residuals as input. The remaining data included 10 time series of RESP counts and 6 time series of GI counts.

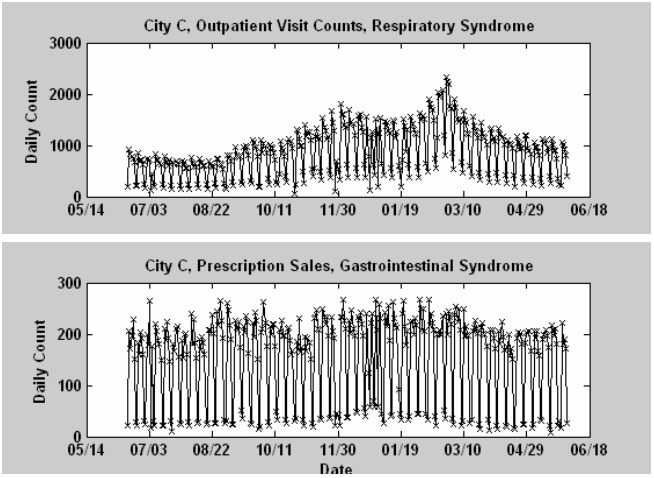

Figure 1. Sample time series of daily counts from respiratory and gastrointestinal syndrome groups.

3. Method

3.1 Data Preconditioning

The data were first adjusted to account for two known and important factors: the day-of-week (DOW) effect and holidays. While syndromic data can come from a variety of sources (store sales, ER admissions, school absences, etc.), these two effects are present in many time series of daily syndromic counts. An exploratory analysis of our data confirmed that the day-of-week and holiday effects are present and strong. For example, many clinics and physician offices have reduced hours on weekends and holidays, so the corresponding visit counts are much lower than those on weekdays.

3.1.1 Day-of-week adjustment

To account for DOW effects we used an adaptive ratio-to-moving-average with multiplicative effects. The ratio-to-moving-average method is similar to the X-11 and X-12 deseasonalizing method employed by the Census Bureau and used in many business applications. The method is based on estimating and removing the trend, and then estimating and removing the seasonal day-of-week effect from the de-trended data. We applied the method to a moving window of 128 days to capture the changing nature of the day-of-week effect. Results from this procedure are shown in Figure 2.

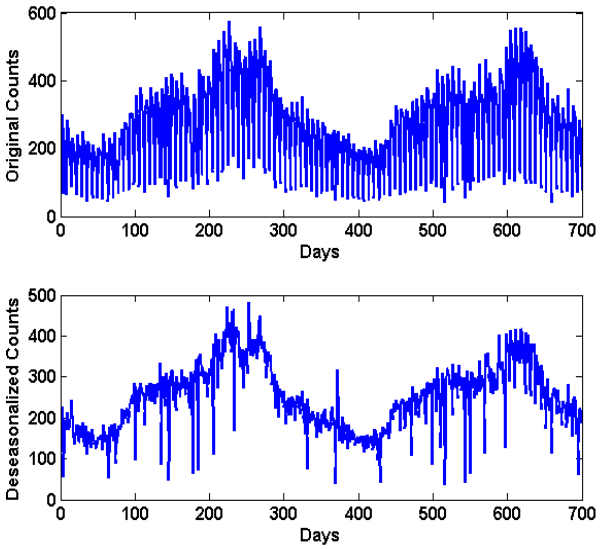

Figure 2. Series before (top) and after (bottom) removing the day-of-week effect.

3.1.2 Holiday adjustment

To account for holidays we first assembled a list of all federal holidays from the Federal Office of Personnel Management (http://www.opm.gov/Fedhol/index.asp). Data points on these days were then replaced with predictions either from the wavelet prediction algorithm

or, when the data history was insufficiently long, with the value from the same day in the previous week.

3.2 Wavelet Decomposition

Wavelet decomposition was performed using a Haar wavelet function. This function is the most basic of the wavelet functions, as it performs a simple averaging and differencing at each level, and was chosen in order to minimize the introduction of edge problems by the algorithm (Shmueli, 2005). The wavelet decomposition is performed as follows: For each level j=1,…,J, a set of approximation and detail coefficients is created (the original time series is considered to be level j=0):

$$Approx_j(i) = \frac{Approx_{j-1}(i) + Approx_{j-1}(i-2^{j-1})}{\sqrt{2}}$$

$$Detail_j(i) = \frac{Approx_{j-1}(i) - Approx_{j-1}(i-2^{j-1})}{\sqrt{2}}$$

The final approximation level ($Approx_J$) together with all detail levels can be used to completely reconstruct the original series. In fact, there is duplicated information; the series may be reconstructed completely if 2^j of the coefficients at each level j are discarded. Discarding these coefficients is called downsampling. If downsampling is performed, then each of the detail levels and the final approximation coefficients are approximately uncorrelated. This correlation removal can be very useful, but in the context of forecasting it makes a wavelet method untenable for prospective prediction or alerting. The reason is that the "holes" mean that at most given time points some of the levels will have deleted coefficients. Instead, we do not downsample, thereby creating a "stationary wavelet transform" (SWT) where we are guaranteed to have coefficients at every level at every time point. This is illustrated for one of the series in Figure 3. The price is that the coefficients are no longer uncorrelated and that the set of time series at each level are interdependent.

An important modification to the ordinary SWT is "shifting into the past" such that coefficients are computed from only present and past data values but not future ones (see Shmueli, 2005 for details).

3.3 Data Prediction Algorithm

The stationary wavelet transform described above was used to perform both 1-day ahead and 7-day ahead predictions. Because the coefficients within a certain level are no longer uncorrelated, we used either an autoregressive (AR) model or a simpler exponentially weighted moving average (EWMA) with smoothing coefficient set to λ=0.3 in order to predict the next coefficient value. These predicted coefficient values at each level were then combined to produce the desired n-days ahead prediction of the original series.

To compare prediction performance between different wavelet techniques, we computed the median absolute percent error (MedAPE) of the residuals using the predictions from the last 350 days of data. The median was used rather than the mean for this performance metric to reduce the impact of outliers on the statistic. The choice of a percentage error allows comparisons across time series.

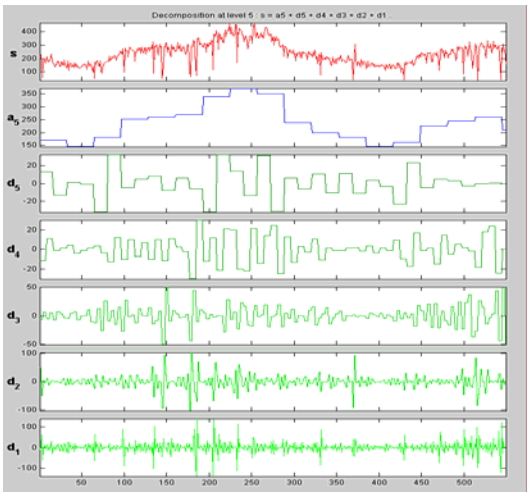

Figure 3. A 5-level Haar Stationary wavelet transform of a preconditioned respiratory syndrome series (top) into an approximation (a5) and five detail levels (d1-d5).

3.4 Alerting Algorithm

To determine an alert, we used a simple prediction-based Shewhart control chart method. The wavelet-based predicted count was compared to the actual count, and the alert level was generated as the difference divided by the standard deviation of the differences excluding holidays from the past 28 days. The threshold for the control chart was varied to generate a receiver operating characteristics (ROC) curve to examine the probability of detection versus the probability of a false alarm.

3.5 Simulated Outbreaks and Evaluation

Our data do not contain a set of easily identifiable and universally accepted disease outbreaks. Although there was a set of outbreak dates suggested by a set of experts, it was not very reliable (Shmueli, 2005). Therefore, in order to evaluate the performance of our method in the presence of outbreaks, we injected the data with one of two types of outbreaks: a one-day spike and a lognormal curve using mean and variance parameters derived from published observations of the incubation period distribution of human cases of avian flu (WHO, 2005). We applied these parameters because of the public's growing concern about the possibility of a human pandemic, and the recent observations of the global increase of H5N1 infection in bird populations suggest that a related virus could be the source of the next such

pandemic. Given that the symptomatology of the next pandemic strain is unknown, we considered both gradual and explosive signals to simulate the data effects of an outbreak.

The data used for this study started in the middle of one calendar year and continued for 700 days, encompassing one full and two partial calendar years. To evaluate the performance of the algorithm for detection, we ran 365 separate trials, inserting one outbreak per trial starting on a unique day in the full calendar year. The spike outbreak inserted had a magnitude equal to one standard deviation of the previous 28 days of raw data. The detection algorithm's output for all trials was then combined and used to generate a ROC curve.

4. Results

4.1 Optimal Configuration

Numerous configurations for the wavelet algorithm were studied to yield the optimal performance. Variations on the length of the training window (including "infinite", using all previous data), the order of the AR model used for predicting each level's coefficients (or using a simple EWMA), and what forms of preprocessing to use (day-of-week, holiday, or neither) were examined to determine which yielded the best performance when predicting next day and 7-day ahead forecasts. A 128-day sliding window, with a 7-day AR, combined with day-of-week and holiday preconditioning, proved most effective. The final predictions are re-seasonalized using the appropriate day-of-week indexes that were estimated in the initial preprocessing step.

4.2 Background Prediction Accuracy

Table 1 gives the median absolute percentage error (MedAPE) statistics for several different configurations of the wavelet-based predictor when no outbreaks were injected. The left columns (blue) are based on a 5-level HAAR wavelet, with a 128-day sliding window for the 7-day AR and full DOW and holiday preconditioning. The middle columns (green) are based on the same technique but without removing the DOW effect, clearly showing the need to deseasonalize the data before applying the wavelet transform. Similarly, the right columns (orange) are based on the same prediction technique but without accounting for calendar holidays. Interestingly, holidays seem to impact the prediction results more for respiratory count time series (containing DOW and annual variations) than the GI count time series.

Table 1. MedAPE of the residuals for GI and respiratory count series comparing the prediction performance of three variations of the wavelets-based algorithm for both 1 and 7-day ahead predictions

GI Count Series

	1-Day Ahead			***7-Day Ahead***		
City	Wavelet	w/o Deseasonalizing	w/o Holidays	Wavelet	w/o Deseasonalizing	w/o Holidays
(a)	12.17	15.07	12.15	10.5	13.94	10.71
(b)	7.94	11.93	9.66	7.59	11.2	9.44
(c)	8.77	12.9	10.08	9.45	11.69	9.84
(d)	7.54	10.22	8.91	7.28	8.94	7.96
(e)	10.97	13.78	11.14	9.55	11.17	10.56
(f)	13.53	16.33	14.24	12.54	15.59	12.1

Respiratory Count Series

	1-Day Ahead			***7-Day Ahead***		
City	Wavelet	w/o Deseasonalizing	w/o Holidays	Wavelet	w/o Deseasonalizing	w/o Holidays
(g)	8.59	11.31	10.8	9.74	11.16	10.88
(h)	8.51	12.11	10.46	8.97	10.29	11.62
(i)	5.24	7.75	6.65	6.67	7.22	7.76
(j)	10.66	15.61	12.12	12.75	16.73	14.65
(k)	10.15	13.87	11.7	11.99	13.04	13.81
(l)	11.25	15.29	12.81	15.33	14.95	16.94
(m)	13.45	18.17	14.52	13.44	17.58	14.28
(n)	5.96	8.29	6.55	6.9	7.73	7.85
(o)	7.3	10.06	9.86	9.68	10.28	11.49
(p)	8.18	11.39	9.58	11.37	10.16	12.67

4.3 Outbreak Detection

To evaluate the ability of the algorithm to detect an outbreak we injected such patterns into the data. Figure 4 shows initial results for two series, one for respiratory counts (top) and one for GI counts (bottom) where a spike-shaped outbreak was injected on each day in the data. ROC curves are generated by the day-of-week on which the outbreak was injected. Overall, the detection probability is high, but the performance varies such that it is most likely to detect a weekend outbreak and hardest to detect one on a Monday.

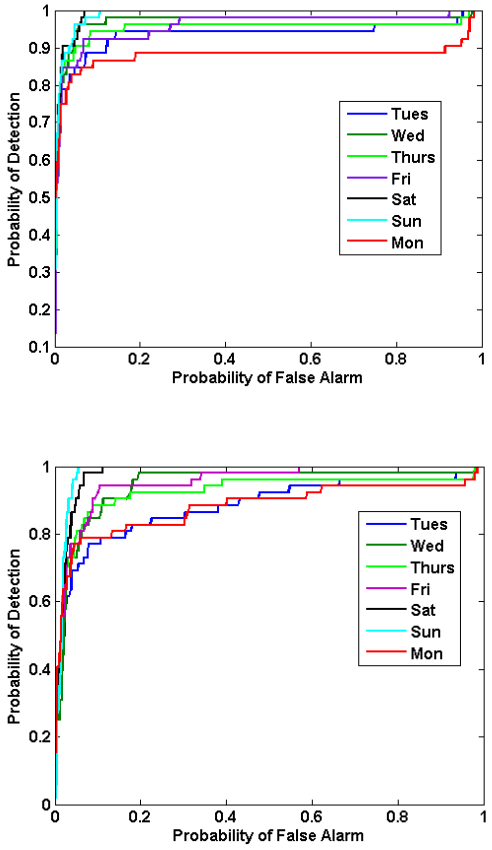

Figure 4. ROC curves for a respiratory series (above) and GI series (below), according to the day of week that a spike-outbreak was injected.

Table 2 shows the sensitivity of the algorithms to spike outbreaks at specific false alarm rates. For example, if one accepts a false alarm every two weeks, then in city (g) the algorithm would be expected to detect 82% of spike outbreaks. These numbers are for spike outbreaks having size equal to one standard deviation of the data. A larger spike would be more likely to be detected.

Table 2. Detection accuracy comparing specific false alarm rates on each series.

Probability of Outbreak Detection for Specific False Alarm Rates		
Gi Count Series		
City	1 False alarm Every 2 weeks	1 False alarm Every month
(a)	0.7269	0.5038
(b)	0.8818	0.6104
(c)	0.8299	0.6466
(d)	0.8682	0.7648
(e)	0.8400	0.7818
(f)	0.6219	0.2638
Respiratory Count Series		
City	1 False alarm Every 2 weeks	1 False alarm Every month
(g)	0.8294	0.7315
(h)	0.8593	0.7509
(i)	0.9534	0.9178
(j)	0.8137	0.6507
(k)	0.8317	0.6650
(l)	0.8667	0.5102
(m)	0.6975	0.4682
(n)	0.9126	0.8438
(o)	0.9096	0.8569
(p)	0.8466	0.7309

5. Discussion

The method described in this paper is based on Goldenberg et al. (2002), but improves upon it in four aspects. First, an important step of preconditioning that directly treats day-of-week and holidays is incorporated instead of the more general smoothing method using a cosine-transform. Second, the wavelet decomposition uses a Haar and "backward shifted" coefficients to minimize edge problems and enable prospective operation rather than retrospective analysis. Third, the method is applied to a wider range of syndromic data from a geographically diverse set of metropolitan areas. And

fourth, we evaluate the performance of the algorithm for detecting two different types of outbreak. This broad testing highlights the method's advantage as a wide-range detector.

From comparisons with other methods, the performance of the wavelet-based detector for univariate time series appears to be comparable to methods such as Holt-Winter's exponential smoothing and adaptive regression models (see Burkom et al. 2006 for further details). However, the distinguishing utility of the wavelet-based methods is likely to be their robustness over many data types and applications to the multivariate problem. Our next steps are to investigate its performance on low-count data, and on a larger variety of outbreak patterns. Another challenge is to include the day-of-week handling within the wavelet detector and eliminating the need for a preconditioning step. This is challenging when using a Haar wavelet, because it is by nature dyadic and therefore "skips" the 7-day periodicity. One direction is to apply the wavelets to an "8-day" week obtained with interpolation.

References

Aha, D. W. (1990). *A study of instance-based algorithms for supervised learning tasks: Mathematical, empirical, and psychological evaluations.* Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine.

Fisher, D. H. (1989). Noise-tolerant conceptual clustering. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 825-830). San Francisco: Morgan Kaufmann.

Burkom, H.S., Murphy, S.P., and Shmueli, G. (2006). Automated Time Series Forecasting for Biosurveillance, *submitted*.

Goldenberg, A., Shmueli, G., Caruana, R.A., and Fienberg, S.E. (2002). Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales, *Proceeding of the National Academy of Sciences*, vol. 99, Issue 8, pp. 5237-5240.

Ryan, T.P. (1989). *Statistical Methods for Quality* Improvement. John Wiley & Sons: NY.

Shmueli, G., (2005). "Wavelet-Based Monitoring in Modern Biosurveillance", *Working Paper, RHS-06-002, Robert H Smith School, University of Maryland* (http://ssrn.com/abstract=902878).

Shmueli, G., and Fienberg, S.E. (2006). "Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance", *Statistical Methods in Counter-Terrorism*, Eds: A Wilson and D Olwell, Springer, forthcoming.

Siegrist, D. and Pavlin, J. (2004). "BioALIRT Biosurveillance Detection Algorithm Evaluation." Syndromic Surveillance: Reports from a National Conference, 2003. *MMWR* 53(Suppl): 152-158.

Stacey, D., Calvert, D., Shu, J., and Harvey, N. (2005). "A Preliminary Wavelet Analysis of OTC Pharmaceutical Sales Data," *Working paper, University of Guelph.*

The Writing Committee of the World Health Organization (WHO) Consultation on Human Influenza A/H5 (2005). "Current Concepts: Avian Influenza A (H5N1) Infection in Humans." *N England Journal of Medicine*, 353:1374-1385.

Zhang et al. (2003). "Detection of Outbreaks from Time Series Data Using Wavelet Transform", *AMIA Fall Symp, Omni Press CD*, pp. 748-752.