Implementation and Comparison of Preprocessing Methods for Biosurveillance Data

Thomas Lotze[1], Sean Murphy[2] and Galit Shmueli[1,3]

[1]Applied Mathematics and Scientific Computation Program, University of Maryland, College Park, MD 20742
[2]The Johns Hopkins University Applied Physics Laboratory
[3]Department of Decision and Information Technologies and the Center for Health Information and Decision Systems, Robert H Smith School of Business, University of Maryland, College Park, MD 20742

Abstract

Modern biosurveillance relies on multiple sources of both pre-diagnostic and diagnostic data, updated daily, to discover disease outbreaks. Intrinsic to this effort are two assumptions: (1) the data being analyzed contain early indicators of a disease outbreak and (2) the outbreaks to be detected are not known a priori. However, in addition to outbreak indicators, syndromic data streams include such factors as day-of-week effects, seasonal effects, autocorrelation, and global trends. These *explainable* factors obscure *unexplained* outbreak events and their presence in the data violates standard control chart assumptions. Monitoring tools such as Shewhart, Cumulative Sum, and Exponentially Weighted Moving Average control charts will alert largely based on these explainable factors instead of on outbreaks. The goal of this paper is twofold: first, to describe a set of tools for identifying explainable patterns such as temporal dependence, and second, to survey and examine several data preconditioning methods that significantly reduce these explainable factors, yielding data better suited for monitoring using the popular control charts.

1 Introduction: Monitoring via Control charts

Control charts are a tool for monitoring a process parameter (such as the process mean) by comparing daily parameter estimates to pre-determined thresholds called *control limits*. While other methods were proposed and sometimes used ([1], [2], [3]), control charts remain one of the most popular monitoring tools in traditional and modern biosurveillance. Such charts are applied in the widely-used monitoring systems BioSense ([4]), RODS ([5]), EARS ([6]), and ESSENCE ([7]). The classic Shewhart chart for monitoring the process mean relies on drawing a sample from the process at some frequency (e.g., weekly), and plotting the sample mean on the chart. Parameter limits are defined such that if the process remains in control, all (or nearly all) of the sample means will fall within the control limits. If a sample mean exceeds the control limits, it is assumed that the process mean has shifted, or in other words, the process has gone out of control; an alarm is triggered and an investigation follows to find its cause(s) ([8], [9]). Figure 1 shows an example of a one-sided Shewhart control chart, for detecting increases in the process mean. The dotted line indicates the control limit; red points show points exceeding the limit.

[Figure 1 approximately here]

To better understand how control charts can be applied to biosurveillance data, we discuss both the original and intended use of these charts. Statistical control charts, invented by Walter Shewhart, were first used in the 1920s to monitor factory outputs to discover abnormally high rates of product defects. An alarm indicated variance beyond the normal operating conditions and the presence of a "special cause", which was usually a faulty process that could then be corrected. Control charts have since been applied to a growing number of areas beyond industrial control, including extensive application to biomedical monitoring [10]. The three most commonly used control charts are: (1) Shewhart charts that monitor values of a sample statistic (e.g., the mean or standard deviation) or individual counts, (2) Cumulative Sum (CuSum) charts that monitor cumulative sums of sample deviations from a target process mean, and (3) exponentially weighted moving average (EWMA) charts that monitor an exponentially weighted average of current and past sample statistics. While all of the different charts monitor deviations from the target value of the process, each one is most effective at detecting a particular type of deviation from the target mean: a single spike, a shift in the process mean, or a gradual increase in the mean, respectively [11]. Underlying all of these methods is the assumption that the monitoring statistics are independent and identically distributed (iid), with the distribution generally assumed normal (although modifications can be made for statistics with known, non-normal distribution). While control charts are very effective for monitoring processes that meet the independence and known distribution assumptions, they are not robust when these assumptions are violated [12]. If the control chart assumptions do not hold, then they will fail to detect special cause variations and/or they will alert frequently even in the absence of special cause variations.

1.1 Challenges with Biosurveillance Data

Modern biosurveillance data come from multiple sources and in many forms. In general, syndromic data tend to be indirect measures of a disease (as opposed to more traditional diagnostic or clinical data). Examples are daily counts of emergency room visits, over-the-counter (OTC) or prescription medication sales, school absences, doctors' office visits, veterinary reports, or other data streams that could contain an indication of a disease outbreak. The purpose of biosurveillance is to monitor these time series to detect disease outbreaks. As in the industrial setting, control charts are used to monitor such data to detect "special causes" or abnormalities that are potentially indicative of an outbreak. However, currently collected biosurveillance data violate most of the assumptions required of data monitored by control charts. Thus, alarms triggered by control charts applied directly to raw syndromic data can arise not from actual outbreaks but due to explainable patterns in the data. Reports of very high false alarm rates from users of current syndromic systems lend evidence to this claim.

The explainable patterns are caused by factors unrelated to a disease. As an example, it is quite common for doctors' offices to have reduced staffing on weekends. Therefore, a syndromic data stream capturing daily doctor visits will see an explainable and predictable drop on Sundays and a corresponding increase on Monday. Many syndromic data streams demonstrate a marked day-of-week (DOW) effect, dropping or increasing in counts over the weekends with an early work-week resurgence or drop. Holidays and other external factors can cause a similar phenomenon. Even the release of Harry Potter books has a known effect on hospital admissions [13].

1.2 Effect of Assumption Violations on Control Charts

Although different data streams exhibit different behavior, a few explainable patterns exist that are common to many series and that clearly violate control chart assumptions. The presence of explainable components in syndromic data leads to a direct violation of the assumption of iid counts. These effects have been also seen in traditional surveillance systems ([14]) but are even more pronounced in syndromic surveillance: due to its more frequent collection, it is subject to greater autocorrelation; due to its data source being less direct, it is more influenced by components not related to disease outbreak. Therefore, in order to increase the effectiveness of control chart monitoring for biosurveillance, we must account for these components in the raw data. Optimally, once these components are removed, we would be left with an iid series. Any shifts in the process would then be attributable to unexplained components; the likelihood that a shift in

the process corresponds to a disease outbreak would then increase, thereby increasing the probability of detection while decreasing the probability of a false alarm.

An example of this can be seen by comparing a CuSum chart applied to the series of daily sales of allergy medication, before (Figure 2, top) and after (Figure 2, bottom) preconditioning.

[Figure 2 approximately here]

Here, the preconditioning significantly reduces the impact of seasonality. It is important to note that it is not simply the number of alerts that decreases due to the preconditioning (this could be achieved by simply raising the alerting threshold), but that the pattern of the alerts changes. Instead of multiple alerts and generally high levels for each season, we see that there are much tighter spikes after preconditioning. The overall level is much lower and the seasonal impact is reduced, indicating that other deviations (of more interest) would be more easily noticed in the preconditioned data. This can be achieved by removing explainable patterns from the data. We now describe the most prominent explainable patterns found in biosurveillance data, and their effect on chart assumptions.

The first explainable pattern is cyclic behavior including day-of-week and seasonality: the magnitudes of syndromic data often vary widely as a function of the day-of-week or time of year. If left uncorrected, this source of variation inflates the "normal variation" assumed by the control chart, thereby leading to overly conservative control limits; in biosurveillance, this can result in outbreaks being detected late or not at all. Alternatively, if the control limits are set low enough to detect true outbreaks, they will also be low enough to be set off by normal seasonal variance, resulting in much higher false alarm rates.

The second explainable pattern is daily autocorrelation. Because syndromic data typically arrive daily, there is almost always a strong degree of autocorrelation; sequential daily counts are not independent (because so much of the variance for sequential days comes from common causes). This can result in a higher level of false alerts in CuSum and EWMA charts, as one abnormal count is likely to be followed by more abnormal counts.

A third pattern is related to holidays. Holidays strongly impact the public's consumption of health care, drastically impacting (usually decreasing) many syndromic data streams. These outliers artificially increase the sample variance, causing the same issues as seasonal variation. In addition, many detection methods are sensitive to these "negative singularities" and will falsely report outbreaks when comparing new counts to past low holiday values [2].

Finally, particularly for syndromic series with very low counts (such as the number of unexpected deaths), the distribution of daily counts is far from normal, causing standard control limits to be incorrectly set. Applying control charts directly to raw syndromic data will either fail to detect actual outbreaks or will frequently alarm, despite the absence of actual outbreaks. While a high false alarm rate can be accounted for by artificially widening the control limits, this will then reduce the power to detect true deviations of interest. The reverse is also true. To remedy these problems we propose to remove explainable patterns from the raw data in an attempt to come closer to meeting the assumptions of control chart methods, and thus achieving a better ratio of detection power to false alarm rate.

If these explainable patterns are not removed, the control chart assumptions will not be met, and so the resulting alerts will be largely based on these patterns. Because they can have such a dramatic impact on the quality of the resulting control chart analysis, determining the most effective method for removing these patterns from a given dataset is very important. The tools used in this paper show quantitative and qualitative methods for comparing methods' applicability to a syndromic data series and effectiveness at removing the explainable effects.

The paper proceeds as follows: Section 2 describes the syndromic data used throughout the paper. Section 3 describes a set of statistical tools for detecting explainable patterns that "contaminate" the data, which make direct control chart monitoring ineffective. We then apply these tools to the syndromic data and illustrate their use. The paper's final goal is to survey and evaluate popular methods for data preconditioning. These methods are aimed at removing explainable patterns from the raw data, thereby creating "conditioned data" that can be monitored more effectively using control charts. Section 4 describes different preconditioning methods and evaluates their usefulness by applying them to syndromic data. A comparison is given in Section 5. We conclude and describe future directions in Section 6.

2 Data Description

2.1 Over-the-counter (OTC) medication sales

The first dataset, from a grocery chain in the Pittsburgh area, includes daily sales for seven categories of medications, from August 1999 to January 2001 [15]. Average daily counts vary largely across different categories, with varying degrees of weekly and annual dependence. The seasonal factor dominates the time series, and would therefore be the major cause for alerts in standard control charts. In the following we use one of the series (throat lozenges) to illustrate the behavior of OTC series. Two others appear in the Appendix (Asthmatic and Allergies medications).

2.2 Chief complaints at emergency departments

The second dataset, from ESSENCE (Electronic Surveillance System for the Early Notification of Community-Based Epidemics), is composed of 35 time series representing daily counts of ICD-9 codes. ICD-9 codes are the 9th edition of the International Statistical Classification of Disease and Related Health Problems, published by WHO and used worldwide. These ICD-9 codes are generated by patient arrivals at emergency departments (ED) in an unspecified metropolitan region from Feb-28-1994 to Dec-30-1997. The 35 series were then grouped into 13, using the CDC's syndrome groupings. These syndrome groups show a diversity in the level of daily counts and in weekly and annual dependence across the different syndrome subgroups. The counts for the 38 holidays contained in the dataset were eliminated. In the following we use one series (Gastrointestinal (GI)-related ED visits), and two additional ED visits series (Respiratory and Unexplained Deaths) are displayed in the Appendix.

3 Tools for Detecting Explainable Patterns

There are many tools available for detecting explainable patterns in the data. Although some of these (notably domain knowledge and graph analysis) will require human intervention, this analysis need not be carried out every day; once a series has been analyzed, the preprocessing method can be chosen and applied continuously with occasional checks.Although it is tempting to completely automate the analysis and preprocessing of syndromic data series, human intervention is still a very valuable tool for finding and removing explainable patterns in the data. We now describe each of three methods for detecting explainable patterns.

3.1 Domain Knowledge

The first method for determining temporal patterns in syndromic data is to use domain knowledge. From public health and medical sources we can learn whether there exist day-of-week effects in counts of emergency department visits and doctors' office visits. Since many hospitals dramatically reduce staffing on weekends ([16] [17] [18] [19] [20] [21] [22]), counts are generally much lower on weekends. We also know that seasonal trends might be present in ED visits for reasons such as flu season. Marketing knowledge can tell us that grocery shopping is more popular on weekends than on weekdays. And for both types, holidays always have exceedingly low counts.

3.2 Graphs and charts

The second step is to use statistical summaries and graphs to quantify such effects, and to identify others. Some useful statistics and graphs are:

Time plots of each series with zoomed-in views (for detecting local effects such as DOW).

Moving average charts for detecting overall trends, with narrow windows for local trends and wider ones for global trends. A window of width 7 suppresses DOW effects, whereas width 28 suppresses (nearly) monthly effects.

Moving standard deviation charts for determining whether the seasonal variation is additive or multiplicative (that is, if higher values also lead to a greater standard deviation, as is normally the case in count data). This can suggest using a logarithmic transform or a multiplicative seasonality model.

Autocorrelograms - plots of autocorrelations and partial autocorrelations at different lags for highlighting periodic effects and temporal dependence. A lag 1 correlation indicates daily autocorrelation, a lag 7 and its multiples indicate a day-of-week effect, and a lag 365 indicates a yearly pattern.

Normal probability plots and histograms for assessing normality. Skewness and kurtosis statistics are also useful for this purpose.

Two additional potentially useful statistics and graphs are partial autocorrelations at lag 365 and spectral plots. However, we do not use partial autocorrelograms because partial autocorrelation values for a 365-day lag are very sensitive to missing values, and are not reliable when holidays have been removed. We also do not use spectral plots because they tend to mask weekly seasonality when there is a strong yearly seasonality (the size of the weekly peak can be so small as to be barely distinguishable).

3.3 Summary statistics

The following statistics are useful for detecting patterns and evaluating and comparing preprocessing methods:

mean: the sample mean

stdev: the sample standard deviation

weekendMean: the mean from weekends only; deviations from the global mean are indicative of the magnitude of the weekday-weekend effect

percentInMin: the percent of values which are at the minimum value for the series; higher values indicate that this is a "low count" series

pacfWeek: the partial autocorrelation function (*pacf*) coefficient at 7-day lag; greater absolute values indicate stronger day-of-week effect

acfWeek: the autocorrelation function (*acf*) coefficient at 7-day lag; greater absolute values indicate stronger day-of-week effect. If there is no correspondingly large pacf, it may be indicative of short-term (shorter than 7-day) clustering effects in the series

acfYear: the maximum *acf* coefficient at 364- or 365-day lag; greater absolute values indicate stronger yearly seasonality (364 is preferable because it will be the same day-of-week and often shows a greater degree of correlation than 365)

daysHighPacf: the number of days the series has a "significantly high" (greater than $4/\sqrt{n}$) *pacf*; This measure is often used for determining the autocorrelation order in an ARIMA model, and here can indicate the length of significant local (not yearly) seasonality effects

skewness: the skewness of the series; deviations from 0 indicate non-normality, with positive deviations indicating a positive skew and negative deviations indicating negative skew

excessKurtosis: the kurtosis-3; deviations from 0 indicate non-normality, with positive values indicating more centrally peaked data and negative values indicating larger-tailed data

3.4 Applying graphs and summaries to data

A close examination of the characteristic plots and summary statistics can be used to detect different explainable patterns in the data. The top row in each of Figures 3 and 4 present several of these plots for sales of OTC throat lozenges (Figure 3) and GI-related ED visits series (Figure 4). Corresponding summary statistics are given in the left column of Tables 1 and 2. See the Appendix for plots and statistics for four additional series.

Seasonality

The degree of seasonality in the data can often be determined by a visual inspection of the time series at different temporal scales. Autocorrelograms (ACF and PACF plots) and spectral plots are two additional plots useful in uncovering seasonal patterns. However, one must be careful in using spectral plots, as large peaks can mask seasonal variance at other scales.

Most of the series in our datasets exhibit a pronounced seasonal pattern with peaks during the winter months. This can be seen for two of the series in the top left panels in Figures 3 and 4, where throat lozenge sales exhibit a 3-month cycle and GI-related ED visits exhibit a yearly pattern (see Appendix for similar plots for four more series). This can also be seen in the relative values of the 365-day ACF (sixth column), where a large value indicates a strong yearly seasonal component.

Day-of-week (DOW) effect

To detect day-of-week effects we first zoom-in to a shorter one-month period of the data. The second column (top row) in Figure 3 displays this for

throat lozenge sales and in Figure 4 for and GI-related ED visits. Weekends are highlighted in these zoom plots. Another useful tool are ACF and PACF plots (columns 5-7 in these Figures). Both series exhibit a strong DOW effect: OTC sales have peaks on weekends (due to the general trend of high volume purchasing on weekends), and ED visits drop on weekends. The ACF plot for ED visits and some of the OTC medications (see Appendix for plots of four additional series) shows high autocorrelation at lags 7, 14, 21, etc., indicative of a DOW effect. The DOW effect is even present in unexplained deaths!

Autocorrelation

Autocorrelograms (graphs of the estimated autocorrelation as a function of the lag) are useful for studying the correlation of the data series with itself at various lags, and can indicate lags that play an important role. As mentioned above, the autocorrelograms (columns 5-6 in Figures 3 and 4 and the Appendix) show high autocorrelation at lags 7, 14, 21, etc. for most of the series, indicative of a DOW effect. This can also be seen in the higher values for the acfWeek and pacfWeek statistics (left column in Tables 1 and 2 and the corresponding tables in the Appendix). In addition, examining longer lags indicates bi-annual seasonality for throat lozenges sales, while for GI-related ED counts the weekly pattern repeats and is stronger than a yearly pattern.

Normality

To evaluate how closely the data follow a normal distribution, we use histograms and a normal probability plot and also examine summary statistics such as skewness and excess kurtosis. From the normal probability plots in column 4 of Figures 3 and 4 we see that both series exhibit significant deviations from normality. Throat lozenges (and other OTC sales) are skewed to the right; they also seem to be slightly more tightly clustered than normal data. GI-related (and other) ED counts appear to be bimodal (one peak for weekends, one for weekdays). The deviation from normality is also seen in the values for skewness and kurtosis in the summary tables (Tables 1 and 2)

[Figure 3 approximately here]

[Figure 4 approximately here]

[Table 1 approximately here]

[Table 2 approximately here]

4 Methods for Preconditioning

Several methods exist for removing explainable factors from the data. These include model-based methods, which assume a particular model and estimate the parameters in that model, and data-driven methods, which fit the data non-parametrically rather than attempting to model the causes. The methods can also differ in their global versus local nature.

4.1 Linear regression models

Regression models are a popular method for capturing recurring patterns such as day-of-week, seasonality, and trends [23]. The classic assumption is that these patterns do not change over time, and therefore the entire data can be used to estimate them. To model the different patterns, suitable predictors are created:

Day-of-week effects can be captured by six dummy variables, each representing one day of the week (relative to the remaining baseline day). If there is only a weekday/weekend effect, a single dummy variable can be used.

A global linear trend can be modeled using a predictor t that is a running index ($t = 1, 2, 3 \ldots$). Other types of trends such as exponential and quadratic trends can also be captured via a linear model by transforming the response and/or index predictor, or by adding transformations of the index predictor (such as adding t^2 to capture a quadratic trend).

Seasonality can be modeled by a sinusoidal trend. The Centers for Disease Control and Prevention (CDC) use a regression model that includes sine and cosine functions to capture a cyclical trend of mortality rates due to influenza [24, 25], although these terms will not be significant in series without pronounced seasonality. Another regression-based method for dealing with seasonality is to fit local regression models, using past data from the same

time of year ([26]). Note that explicit modeling of seasonal variation assumes that the seasonal pattern remains constant from year to year.

Holidays can be captured by constructing a dummy variable for holidays or by replacing holiday values with missing values.

From our experience as well as other reports in the literature [27, 28], we find that seasonality effects tend to be multiplicative rather than additive with respect to the response variable. Thus, a linear model where the response is transformed into a natural log ($\log(y)$) is often appropriate. For our data series, we fit a linear regression and a multiplicative regression, and found that the multiplicative version better captured the day-of-week effect. Both are reported below.

Currently, several biosurveillance systems implement some variation of a regression preconditioning. ESSENCE uses a linear regression model that includes day-of-week, holiday, and post-holiday indicators ([7]) and BioSense uses a Poisson regression with predictors that include a linear trend, sine and cosine effects for seasonality, month indicators, DOW indicators and Holiday and day-after holiday indicators ([29]).

Regression models can also be used to integrate external information that can assist in removing explainable patterns. For example, the seasonal pattern was highly correlated with temperature. Figure 5, which shows the relationship between counts of throat lozenge sales (in black) and the average daily temperature (in red), demonstrates this relationship. There is a strong negative relationship between temperature and sales: as the weather gets colder, more cough remedy drugs are sold. However, the causality of temperature is unclear and we therefore treat it only as a proxy. One way is to create alternative temperature-related predictors for capturing yearly seasonality. An example is a *Date function*, which is a linear function rising to 1 in winter and decreasing to -1 in summer.

[Figure 5 approximately here]

The regression model for our data includes daily dummy variables (*Monday, Tuesday, Thursday, Friday, Saturday, Sunday*) to account for the DOW effect, a holiday indicator (*Holiday*), an index variable (*index*) to capture a linear trend, and daily average temperatures (*Tavg*) and monthly dummy variables (*Jan, Feb, Mar, Apr, May, Jul, Aug, Sep, Oct, Nov, Dec*) to remove seasonality. Figure 6 shows an example of the resulting time series of residuals (actual value - predicted value) for the sales of throat lozenges. This series was later used as input into standard control charts.

[Figure 6 approximately here]

The main advantage of regression modeling is that it provides a general yet powerful method to remove variation due to factors unrelated to outbreaks. It is relatively effective at removing both yearly seasonality and day-of-week variation. However, it requires a fairly large amount of data for obtaining accurate estimates, especially for long-term patterns.

4.2 Ratio-to-moving-average indexes

For cyclical data, with virtually any cycle length (weekly, monthly, yearly, etc.), we can compute seasonal indexes and use them to deseasonalize the data. Seasonal-adjustment methods are very popular in business and government agencies such as the Bureau of Labor Statistics and the Census Bureau use such methods to report figures such as monthly unemployment rates.

A simple method to compute indexes is the ratio-to-moving-average method. This is also the basis for the X-11 and X-12 systems used by the Census Bureau [30]. The idea is to estimate and remove any linear trend from the data, and then to estimate the seasonal component in the de-trended data. To compute day-of-week seasonal indexes the following algorithm is used:

1. Estimating the trend: For each day, compute the moving average with a 7-day window centered around that day. For example, for a Tuesday and a window of seven days, we compute the average of the 3 previous days (Sat, Sun, Mon), the value on Tuesday itself, and on the 3 following days (Wed, Thur, Fri).

2. Removing the trend: Divide the daily value by its corresponding moving average. These are the *raw seasonals*.

3. Estimating seasonality: Compute the average of all raw seasonals for the same day (e.g, the

raw seasonal for each of the Tuesdays are averaged across the entire period).

4. Scale the averages so that they sum to 1.

This algorithm gives multiplicative indexes, where each index gives the percentage of counts on that day relative to the weekly average. For example, an index of 1.2 for Tuesday would mean that Tuesdays have 120% higher counts than the average weekly count. A similar process can be followed to compute monthly indexes or any other fixed period.

It is possible to compute and remove multiple seasonal cycles with different periods such as a weekly cycle and an annual cycle. For the syndromic data we tried both a 7-day seasonal adjustment (DOW effect) and a yearly adjustment (done using approximate monthly seasons: 365-day period, 12 seasons), as well as combinations of the two (first weekly adjustments, then yearly; and vice-versa). While the 7-day procedure is quite effective at removing weekly patterns, it obviously does not remove yearly seasonality. The yearly deseasonalizaion technique is somewhat effective at removing day-of-week and seasonal patterns, but less so than the other methods described in this section. Seasonal adjustments via ratio-to-moving-average indexes should only be performed on the raw count data; if this method is performed on normalized data centered around zero (such as regression residuals), it frequently generates abnormally high results due to division by an average very close to zero, thereby creating highly unrepresentative results.

4.3 Differencing

Differencing is the operation of subtracting a previous value from a current one. The order of differencing gives the vicinity between the two values: an order 1 differencing means that we take differences between consecutive days ($y_t - y_{t-1}$), whereas an order 7 differencing means subtracting the value of the same day last week ($y_t - y_{t-7}$). This is a popular method in time series analysis, where the goal is to bring a non-stationary time series closer to stationarity [31]. Differencing has an effect both on removing linear trends as well as removing recurring cyclic components. In the context of syndromic data, the only instance where differencing was suggested is in [32]. They show that a 7-day differencing can be effective at normalizing syndromic data.

In our data, the DOW effect is relatively stable throughout the entire period. We therefore use an order 7 difference. The preconditioned time series is simply the difference between the value on the current day and the value 7 days ago. In addition, we accounted for holidays by removing the values on holidays, and then obtaining differenced values for the 7th day following a holiday by differencing at lag 14 (i.e., subtracting the value from two weeks prior). This improves the method by removing outliers from known (holiday) causes.

The main advantage of differencing is that it is easy and computationally cheap to perform, and so provides an excellent basis for comparison. It is very effective at removing both weekly and monthly patterns but can result in abnormally high results after abnormally low points in the original data (called "negative singularities" by [2]). Another side-effect of seven-day differencing is that it creates strong weekly partial autocorrelation effects and can increase the variance in the data if there is little or no existing DOW effect.

4.4 Holt-Winter's exponential smoothing

The Holt-Winters' exponential smoothing technique is a form of smoothing in which a time series at time t is assumed to consist of three components: a level term L_t, a trend term T_t, and a seasonality term S_t. The k-step ahead forecast is given by

$$\hat{y}_{t+k} = (L_t + kT_t)S_{t+k-M}, \tag{1}$$

where M is the number of seasons in a cycle (e.g., for a weekly periodicity $M = 7$). The three components L_t, T_t, and S_t are updated, as new data arrive, as follows:

$$L_t = \alpha\frac{Y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2}$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \tag{3}$$

$$S_t = \gamma\frac{Y_t}{L_t} + (1-\gamma)(S_{t-M}), \tag{4}$$

where α, β, and γ are smoothing constants that take values in $(0, 1)$. Each component is updated at every time step, based on the actual value at time t.

For our data we use the multiplicative seasonality version because the seasonal effects in our syndromic time series are generally proportional to the level L_t. An additive formulation is also available [33, 34].

The principal advantage of this technique is that it is data-driven and highly automatable. The user need only to specify the cycle of the seasonal pattern (e.g., weekly), and the three smoothing parameters. The choice of smoothing parameters depends on the nature of the data and the degree to which the patterns are local versus global. A study by [28] considered a variety of city-level time series, both with and without seasonal effects. They recommend using the smoothing coefficients $\alpha = 0.4, \beta = 0$, and $\gamma = 0.15$ for seasonal series and $\alpha = 0.1, \beta = 0, \gamma = 0.15$ for non-seasonal series. Following this guideline, we used the first settings for each series that exhibited a one-year autocorrelation higher than 0.15, and the second setting otherwise. In addition, we applied the modification suggested in [28], which does not update the parameters if the actual value deviates from the prediction by more than 50% (to avoid the influence of outliers).

The Holt-Winters method is very effective at capturing yearly seasonality and weekly patterns. Although it is not straightforward to tune the smoothing parameters, the settings provided here proved generally effective for our syndromic data. One point of caution should be made. As in any method that produces one-step-ahead predictions, a gradually increasing outbreak is likely to get incorporated into the background noise, thereby masking the outbreak signal. One solution is to generate and monitor k-day ahead predictions ($k > 1$) in addition to one-day-ahead predictions.

5 Method Comparison

We compare the effectiveness of different methods by examining the preconditioned series for explainable patterns and evaluating their conformity to the iid normal assumption. In particular, we evaluate the degree of seasonality (weekly and yearly) in the data by examining the average autocorrelation and partial autocorrelation values at one week and one year lags. Normality is evaluated by examining the skewness and excess kurtosis, as well as histograms and normal probability plots. While numerous methods and combinations and variations of methods were examined, results from only five methods are presented. These five were chosen based on the ability to reduce explainable effects (and resulting false alarms in standard control charts). The methods are: residuals from regression on the counts, residuals from a linear regression on log(counts), 7-day differencing, 7-day differencing modified for holidays, and forecast errors from Holt-Winters exponential smoothing. Figure 3 compares a the preconditioned lozenge sales series across the five methods using the proposed graphs from Section 4 as well as a CuSum chart. A similar comparison is given for the GI-related ED visits in Figure 4. Summary statistics and charts for these series are in tables 1 and 2. Statistics and figures for additional preconditioned OTC series and ED series are given in the Appendix.

Overall, all five methods greatly improved the data quality for syndromic surveillance via control charts, and therefore each method seems suitable as a first preprocessing step. In particular, we find that differencing, regression, and Holt-Winter's smoothing each significantly reduces seasonal patterns; by examining the CuSum charts, we also see that this results in a narrower, more sharply peaked set of alerts. These graphs also show the effect of preconditioning on the autocorrelation and on normality. Autocorrelation at lags of seven (and its multiples) are greatly reduced, as are very-large lag autocorrelations. The Holt-Winter's smoothing appears to best remove the daily autocorrelation, while regression modeling retains some of this autocorrelation. Multiplicative regression models are better than additive models, especially when seasonality is present. And finally, seven-day differencing appears to perform reasonably well, except for creating large negative lag-7 autocorrelations and partial-autocorrelations. Including the holiday correction does remove these negative partial autocorrelations, indicating that holidays do require special treatment in the preprocessing step. When moving from high-count to low-count series (e.g., UnexplainedDeath), we find that Holt-Winter's smoothing becomes inferior to other methods, and is not able to capture the cyclic patterns well (see Appendix).

Although seasonality can be relatively well accounted for, there are still difficulties in creating normally distributed residuals for some data series; mainly, this seems to be due to a strong, centrally peaked distribution, as indicated in the histograms and the high excess kurtosis. This is especially true for low-count series. However, compared to the raw data, there is definitely improvement in eliminating multiple modalities and in getting closer to normality.

6 Conclusions, Limitations, and Future Work

This paper emphasizes the need to account for explainable patterns in biosurveillance data before applying the widely-used control charts. We present several well-known methods for removing such effects and compare their usefulness. Although we focus here on data that is used in temporal monitoring using control charts, such preprocessing can also be helpful in spatial and spatio-temporal monitoring, when an underlying *iid* assumption exists, such as in the widely-used spatio-temporal scan statistic ([35]).

One future direction is to create an automated application that uses these preconditioning methods to explore and categorize each data series, providing recommendations and rationales for various methods to the end user. This automated expert system could help practitioners determine the methods which would best precondition their data, while allowing them to include domain knowledge. Such a system could perform this function by analyzing the statistics above, selecting appropriate preconditioning methods, and then displaying graphical plots to illustrate the reasons for the each suggested method. The user would then be able to assess which patterns are reasonable in a particular dataset, and based on the system's output, to choose the preferred preconditioning operation(s).

This paper provides a general framework for data preconditioning, but there are several improvements that can follow. First, the tools and methods described here are all aimed at univariate series, where each syndromic series is considered separately. Future work should consider the related multivariate nature of the series, both in choosing preprocessing methods and in the analysis of the preprocessed results. Second, in this work we followed the standard CDC grouping for syndromes to arrive at the 13 ED series and the grocery chain's grouping of OTC remedies. However, it might be useful to examine alternative groupings. For example, grouping the OTC sales data into two categories (headache and cough) or removing all category 2 counts (described by the CDC as "codes that might normally be placed in the syndrome group, but daily volume could overwhelm or otherwise detract from the signal generated") from the ED groups. A third issue is of data quality. For our data, we removed one series from consideration because it contained a change in data collection midway through the period being examined, when additional products were added to the category. Measurement issues such as this, which also include sudden drops in products or delays in provided information, are common in biosurveillance data; it is possible that some of these methods could be used to detect or correct such issues.

In our data, we assumed that there were no known outbreaks. However, it is obvious that the data contain seasons of influenza which affect both ED visits and OTC sales. The problem of unlabeled data in the sense that we do not know exactly when a disease outbreak is present and when there is no disease is a serious one for both modelling and performance evaluation. Our suggestion is to use a time period that is assumed to be outbreak free for the preconditioning step, or at least to suppress suspicious periods from affecting the estimation. A related issue that arises in monitoring daily data is that of gradual outbreaks. Autocorrelation between days (in particular, 1-day autocorrelation) should also be examined and controlled for, in order to approach the statistical independence assumption required for standard control charts. However, a gradual outbreak will also increase the autocorrelation between days (as a rising number of people will show symptoms). It is therefore important to remember the danger of embedding the outbreak signal into the background data. As proposed earlier, one solution is to examine predictions that are farther into the future, and also to use a "guard band" that avoids the use of the last few days in the detection algorithm (similar to the implementation by [36]).

Several additional issues exist when removing explainable effects. Seasonal difference in variance is common in series with seasonal effects; this fluctuation in deviation also causes the series to deviate from an *iid* sample, causing more alerts in high-variance periods than low-variance periods. Low-count or sparse data are also not explicitly examined by this investigation. Numerous detection algorithms have problems handling series with sparse counts and numerous zeroes. A metric comparing preconditioning methods' treatment of low-count series would be a useful contribution. Finally, we note again that syndromic counts on holidays are dramatically different from other days. Since the number of holidays is too small to incorporate into our preconditioning methods, we suggest explicitly labeling them and removing them from consideration, in order to improve reliability. We strongly urge the use

of some mechanism to account for holidays, as they are an explainable cause of significant variation.

From a theoretic detection perspective, the principal concern with any preconditioning technique must be its effect on the sensitivity and timeliness of alerting mechanisms. While proper data preconditioning should result in fewer false alerts due to non-outbreak variation, and a higher probability of detecting actual outbreaks, improper preconditioning can result in unexpected and even unreliable performance. It is therefore essential to study the effects of preconditioning using different methods and to assess the robustness of the results.

From an operational perspective, the main concern with data preconditioning techniques will be the presentation of alerts and the data streams causing the alerts to the end user. A health monitor using an electronic syndromic surveillance system that receives an alarm will typically want to examine the raw data stream causing the alert. With preconditioning, the alerting data stream will potentially not resemble the original data stream, reducing the end user's belief or faith in the system. This aspect of data visualization and user interface must be addressed from the end user's perspective.

Acknowledgements

We thank the Howard Burkom of the Johns Hopkins University's Applied Physics Laboratory, for making the aggregated ED dataset, previously authorized by ESSENCE data providers for public use at the 2005 Syndromic Surveillance Conference Workshop, available to us.

For the first author, this research was performed under an appointment to the U.S. Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05- 06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

References

1. Riffenburgh R, Cummins K: **A simple and general change-point identifier**. *Statistics in Medicine* 2006, **25(6)**:1067–1077.
2. Zhang J, Tsui F, Wagner M, Hogan W: **Detection of outbreaks from time series data using wavelet transform**. In *AMIA Annual Symposium Proceedings* 2003:748–752.
3. Moore A, Cooper G, Tsui R, Wagner M: **Summary of Biosurveillance-relevant statistical and data mining technologies** 2002.
4. Bradley CA, Rolka H, Walker D, Loonsk J: **BioSense: Implementation of a National Early Event Detection and Situational Awareness System**. *MMWR* 2006, **54(Suppl)**:11–19.
5. *RODS Version 4.2 User Manual (rods.health.pitt.edu/RODS%204%202%20User%20Manual.pdf)*.
6. Hutwagner L, Thompson W, Seeman G, Treadwell T: **The bioterrorism preparedness and response Early Aberration Reporting System (EARS)**. *Journal of Urban Health* 2003, **80 (2) Suppl**:89–96.
7. Marsden-Haug N, Foster V, Gould P, Elbert E, Wang H, Pavlin J: **Code-based Syndromic Surveillance for Influenzalike Illness by International Classification of Diseases, Ninth Revision**. *Emerging Infectious Diseases* 2007, **13(2)**, [http://www.cdc.gov/EID/content/13/2/207.htm].
8. Page ES: **Continuous inspection schemes**. *Biometrika* 1954, **41**:100–115.
9. Reinke WA: **Applicability of Industrial Sampling Techniques to Epidemiologic Investigations: Examination of an Underutilized Resource**. *American Journal of Epidemiology* 1991.
10. Benneyan JC: **Statistical quality control methods in infection control and hospital epidemiology, Part I: Introduction and basic theory**. *Infection Control and Hospital Epidemiology* 1998, **19(3)**:194–214.
11. Box G, Luceno A: *Statistical Control: By Monitoring and Feedback Adjustment*. Wiley-Interscience, 1st edition 1997.
12. Shmueli G, Fienberg SE: *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Springer 2006 chap. Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance, :109–140.
13. Gwilym S, Howard D, Davies N: **Harry Potter casts a spell on accident prone children**. *The British Medical Journal* 2005, **331**:1505 – 1506.
14. Farrington C, Andrews N: *Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance*, Oxford University Press 2004 chap. Outbreak detection: application to infectious disease surveillance.
15. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE: **Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales**. *Proceeding of the National Academy of Sciences* 2002, **99**:5237–5240.
16. Tarnow-Mordi WO, Hau C, Warden A, Shearer AJ: **Hospital mortality in relation to staff workload: a 4-year study in an adult intensive-care unit**. *The Lancet* 2000.
17. Czaplinski C, Diers D: **The effect of staff nursing on length of stay and mortality**. *Medical Care* 1998.

18. C K, PJ G: **Nurse staffing levels and adverse events following surgery in U.S. hospitals**. *Image J Nurs Sch* 1998.

19. Blegen M, Vaughn T: **A multisite study of nurse staffing and patient occurrences**. *Nursing Economics* 1998.

20. Strzalka A, Havens D: **Nursing care quality: comparison of unit-hired, hospital float pool, and agency nurses**. *J Nurs Care Qual* 1996.

21. McCloskey JM: **Nurse staffing and patient outcomes**. *Nursing Outlook* 1998.

22. Archibald L, Manning M, Bell L, Banerjee S, Jarvis W: **Patient density, nurse-to-patient ratio and nosocomial infection risk in a pediatric cardiac intensive care unit**. *Prdiatric Infectious Disease Journal* 1997.

23. Rice JA: *Mathematical Statistics and Data Analysis, Second Edition*. Duxbury Press 1995.

24. Serfling RE: **Methods for current statistical analysis fo excess pneumonia-influenza deaths**. *Public Health Reports* 1963, **78**:494–506.

25. CDC: **CDC Syndromic Surveillance site** 2006, [http://www.cdc.gov/mmwr/pdf/wk/mm54su01.pdf].

26. Farrington C, Andrews N, Beale A, Catchpole M: **A statistical algorithm for the early detection of outbreaks of infectious disease**. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1996, **159**(3):547–563.

27. Brillman JC, Burr T, Forslund D, Joyce E, Picard R, Umland E: **Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance**. *BMC Medical Informatics and Decision Making* 2005, **5:4**:1–14, [http://www.biomedcentral.com/content/pdf/1472-6947-5-4.pdf].

28. Burkom HS, Murphy SP, Shmueli G: **Automated Time Series Forecasting for Biosurveillance**. *Statistics in Medicine* accepted 2007 (available at http://www3interscience wileycom/cgi-bin/abstract/114131913/).

29. Kleinman K, Lazarus R, Platt R: **A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism**. *Am J Epidemiol* 2004, **159**:217–224.

30. Bureau UC: **U.S. Census Bureau X-12-ARIMA site** 2006, [http://www.census.gov/srd/www/x12a/].

31. Brockwell PJ, Davis RA: *Time series: theory and methods, 2nd ed.* Springer, New York. 1987.

32. Muscatello D: **An adjusted cumulative sum for count data with day-of-week effects: application to influenza-like illness**. *Presentation at Syndromic Surveillance Conference* 2004.

33. Chatfield C: **The Holt-Winters Forecasting Procedure**. *Applied Statistics* 1978, **27**:264–279.

34. Holt CC: **Forecasting seasonals and trends by exponentially weighted averages**. Tech. rep., Carnegie Institute of Technology 1957.

35. Kulldorff M: **Prospective time-periodic geographical disease surveillance using a scan statistic**. *Journal of the Royal Statistical Society: Series A* 2001, **164**:61–72.

36. Burkom HS, Elbert Y, Feldman A, Lin J: **Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from ESSENCE**. *Morbidity and Mortality Weekly Report (MMWR)* 2004, **53 (suppl)**:67–73, [http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a16.htm].

Tables

Table 1: Comparison statistics for sales of throat lozenges, before ("raw data") and after preconditioning using different methods

	raw data	regression	log_regress	7dayDiff	7dayDiff_holi	holt-winters
mean	909.12	-2.34e-13	-2.90e-15	17.60	10.15	10.19
stdev	349.65	132.40	0.13	189.49	171.54	116.06
weekendMean	962.88	0.64	0.01	20.68	11.43	25.50
percentInMin	0.01	0.02	0.02	0.02	0.02	0.04
pacfWeek	0.08	-0.01	0.00	-0.22	-0.33	0.07
acfWeek	0.85	0.22	0.21	-0.09	-0.02	-0.02
acfYear	0.28	0.07	0.12	0.05	0.03	0.04
daysHighPacf	6	1	2	8	50	1
skewness	0.25	0.77	-0.18	0.56	-0.17	0.14
excessKurtosis	-0.83	3.58	1.20	3.88	1.94	2.94

Table 2: Comparison statistics for gastrointestinal-related ED visits, before ("raw data") and after preconditioning using different methods

	raw data	regression	log_regress	7dayDiff	7dayDiff_holi	holt-winters
mean	117.25	1.97e-14	3.31e-16	2.31	-0.04	0.96
stdev	65.69	27.42	0.28	36.53	30.76	30.10
weekendMean	30.03	1.28e-14	2.65e-16	-0.03	-0.03	-5.61
percentInMin	0.02	0.02	0.02	0.03	0.03	0.03
pacfWeek	0.79	0.19	0.11	-0.29	-0.38	0.15
acfWeek	0.35	0.28	0.15	-0.00	-0.03	0.05
acfYear	0.65	0.05	0.05	0.29	0.15	0.20
daysHighPacf	35	7	7	7	35	35
skewness	-0.19	-1.01	-2.99	1.34	0.06	-0.55
excessKurtosis	-1.15	5.21	17.43	8.34	5.38	3.68

Figures

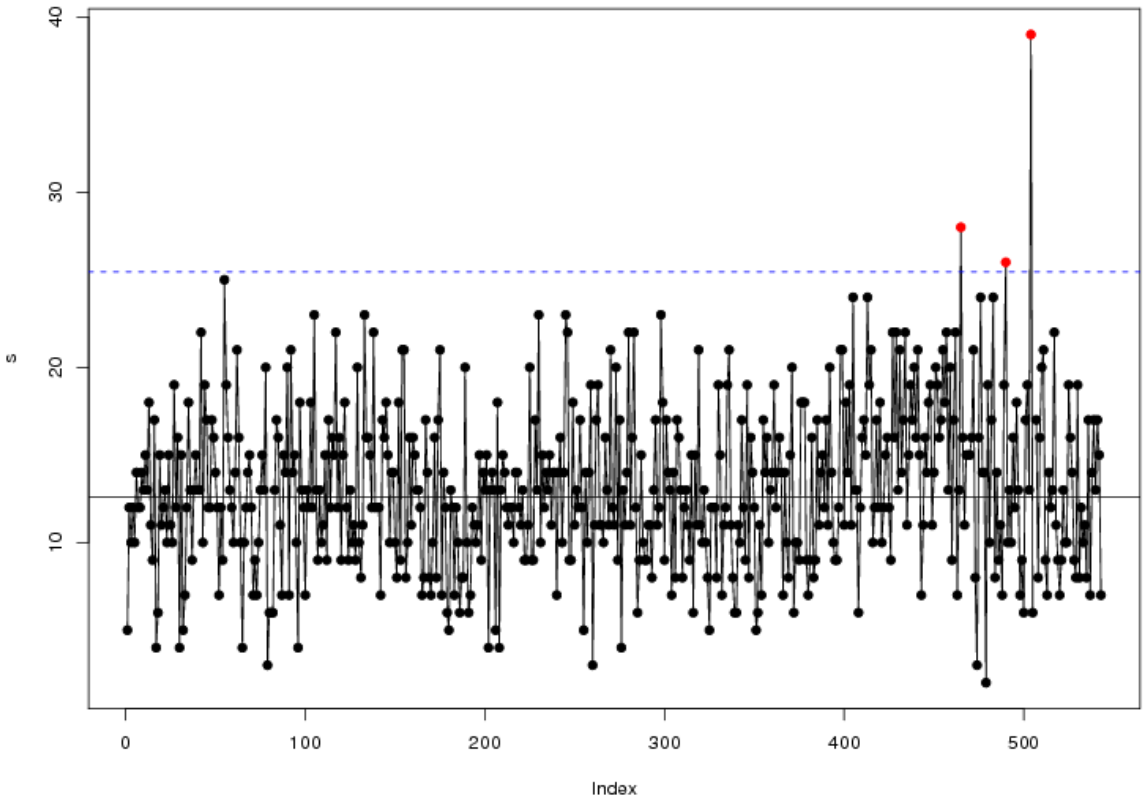
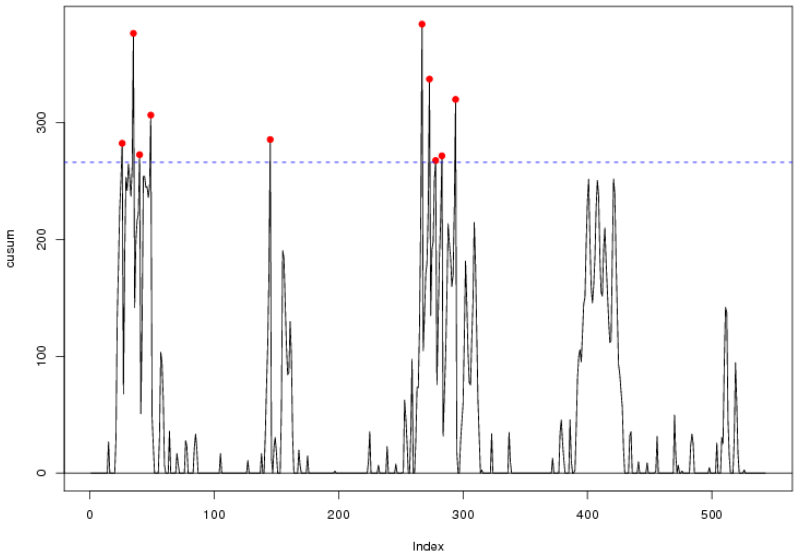


Figure 1: Sample Shewhart Control Chart



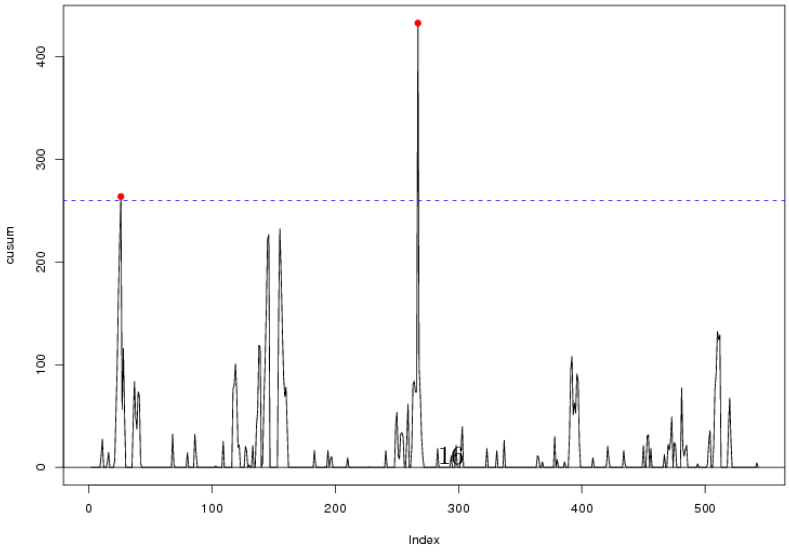


Figure 2: CuSum chart applied to OTC sales data, before (top) and after (bottom) preconditioning

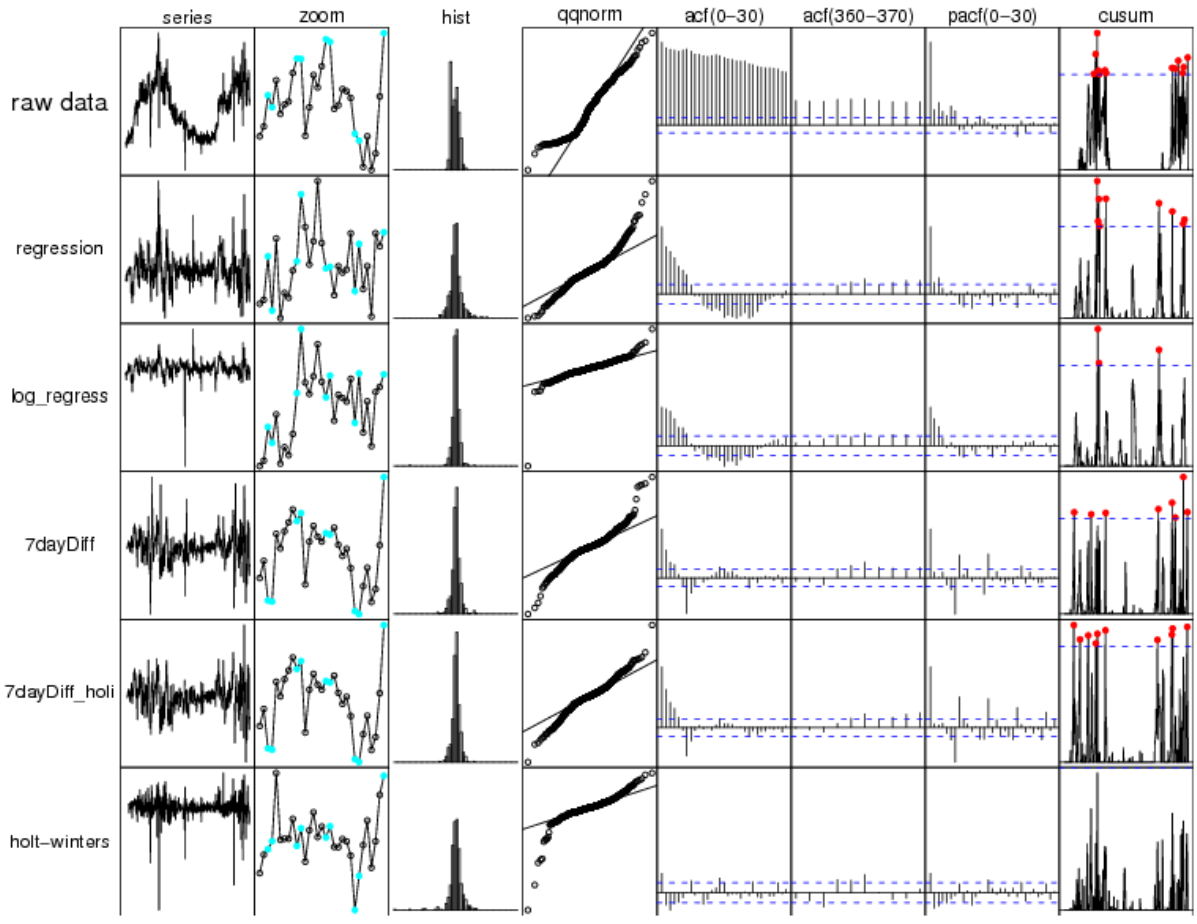


Figure 3: Plots for detecting explainable patterns and comparing preconditioning methods for sales of throat lozenges, for the raw data (top row) and after preconditioning using different methods.

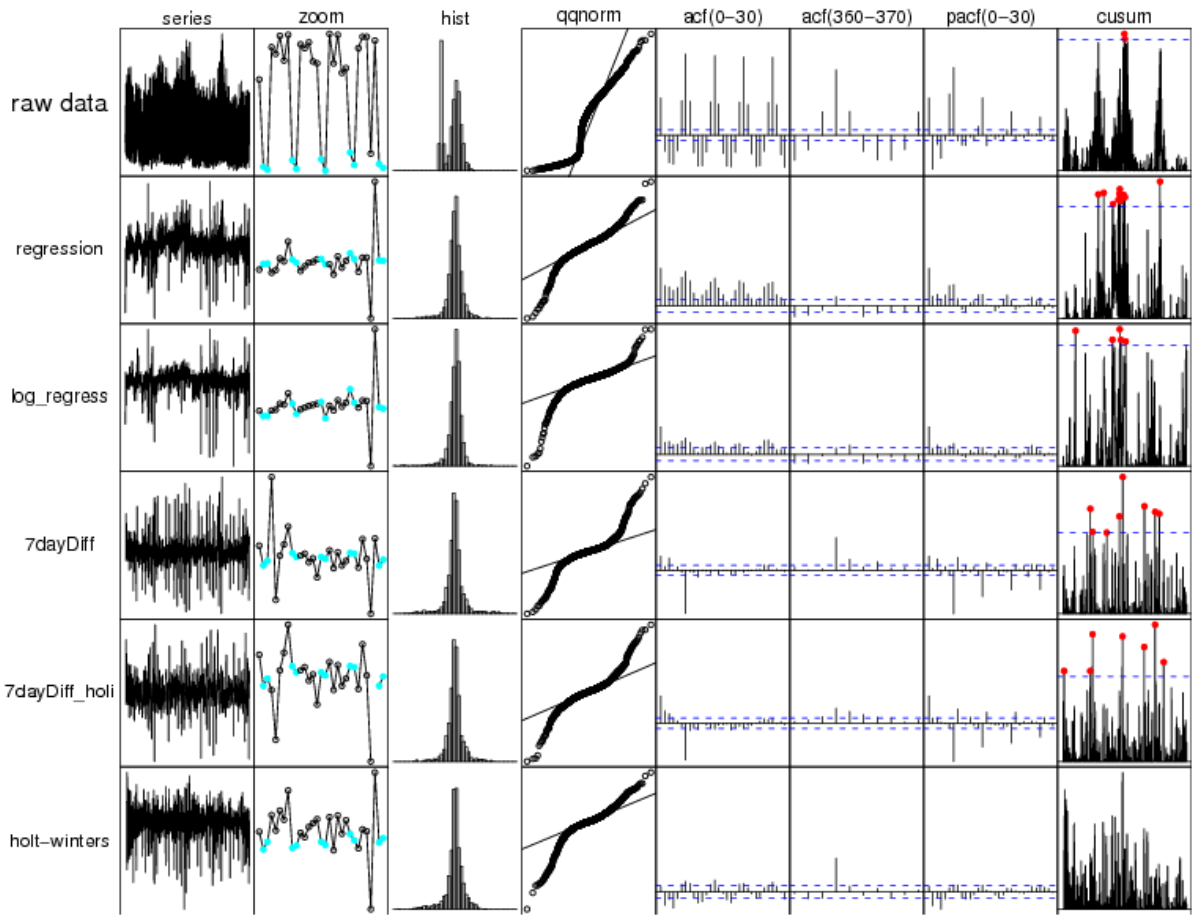


Figure 4: Plots for detecting explainable patterns and comparing preconditioning methods for GI-related ED visits, for the raw data (top row) and after preconditioning using different methods.

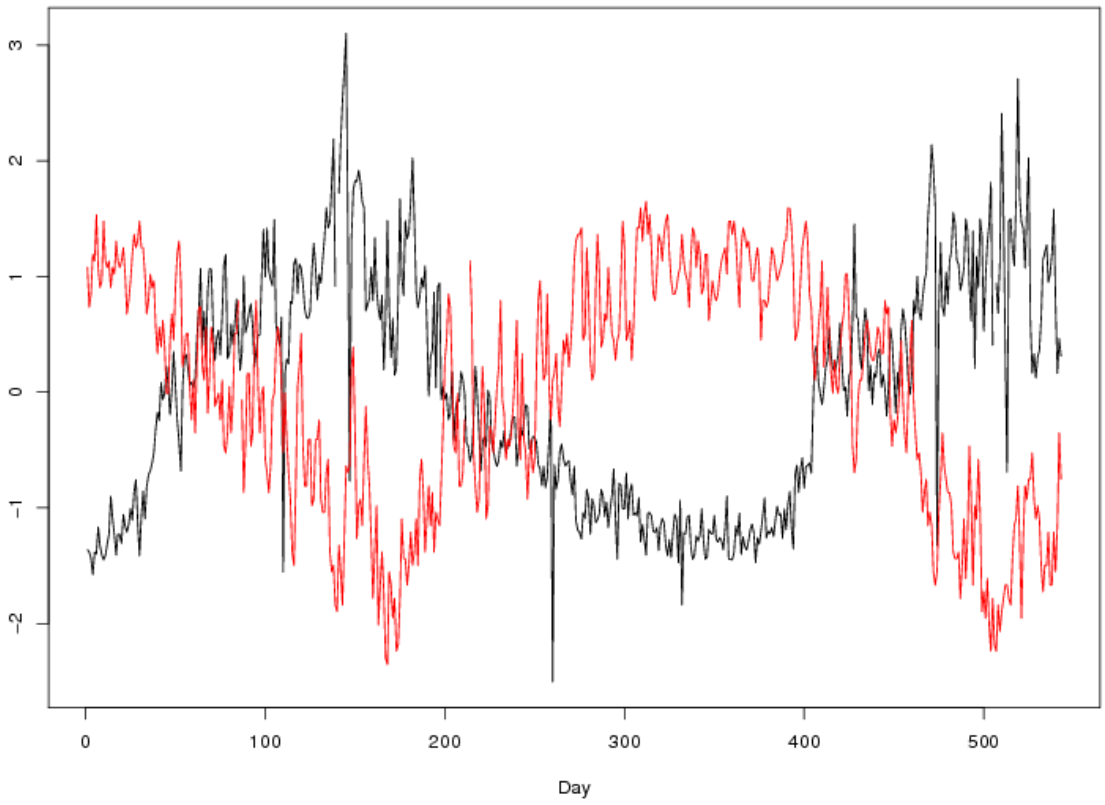


Figure 5: Throat lozenge sales (black) and average temperature (red).

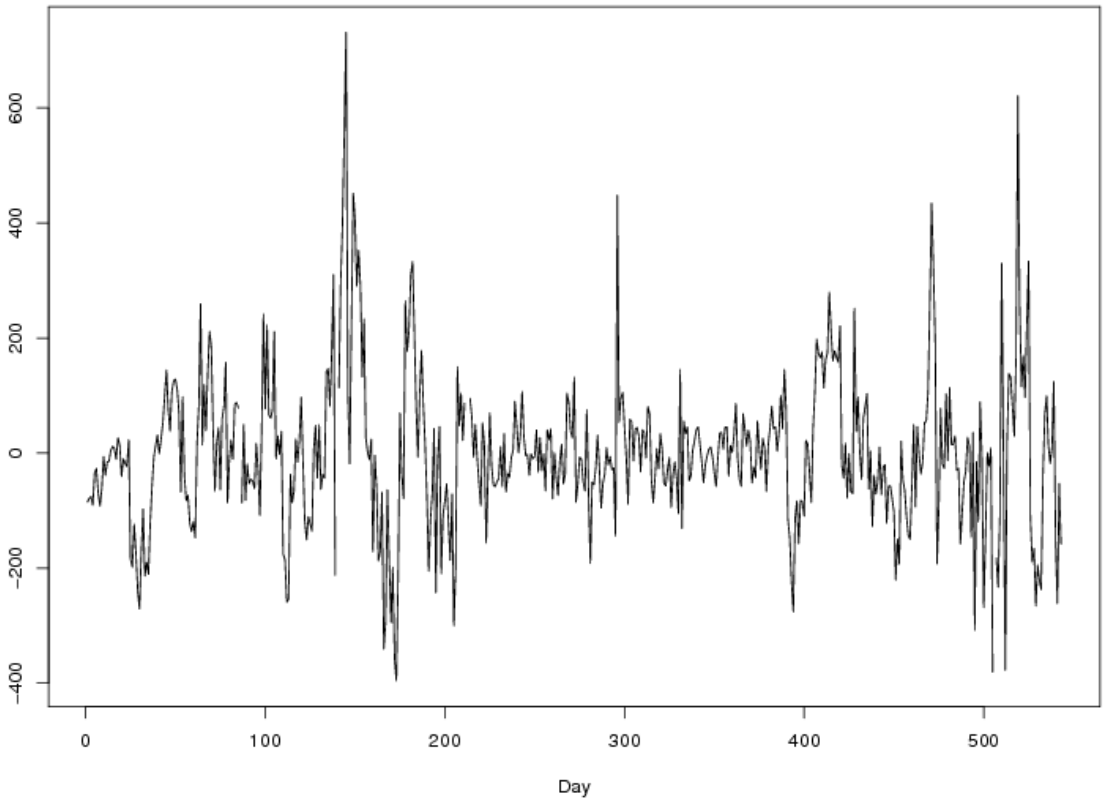


Figure 6: Residuals from a regression model for throat lozenges sales