How does improved forecasting benefit detection?

An application to biosurveillance

Thomas H. Lotze [a] Galit Shmueli [b]

[a] *Applied Mathematics and Scientific Computation Program, University of Maryland, College Park, MD 20742*

[b] *Department of Decision, Operations, and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742*

Abstract

While many methods have been proposed for detecting disease outbreaks from pre-diagnostic data, their performance is usually not well understood. We argue that most existing temporal detection methods for biosurveillance can be characterized as a forecasting component coupled with a monitoring/detection component.

In this paper, we describe the effect of forecast accuracy on detection performance. Quantifying this effect allows one to measure the benefits of improved forecasting and determine when it is worth improving a forecast method's precision at a cost of robustness or simplicity. We quantify the effect of forecast accuracy on detection metrics under different scenarios and investigate the effect when standard assumptions are violated. We illustrate our results by examining performance on authentic biosurveillance data.

Key words: Biosurveillance, Control Charts, Anomaly detection, Sensitivity, Specificity, Timeliness

1 Modern biosurveillance

In modern biosurveillance, time series of pre-diagnostic health data are monitored for the purpose of detecting disease outbreaks. Pre-diagnostic time series typically consist of daily counts of regional emergency department chief complaints such as cough, daily sales of cough remedies at pharmacy or grocery stores, daily counts of school absences, or in general, data that are expected to contain an early signature of a disease outbreak. Outbreaks of interest include terrorist-driven attacks, e.g. a bioterrorist anthrax release, or naturally occurring epidemics, such as an avian influenza outbreak. In either setting, the goal is to alert public officials and create an opportunity for them to respond in a timely manner.

To do this effectively, alerts must occur quickly after the outbreak begins, should detect most outbreaks, and have a low false alert rate. There are a

[1] This work was partially supported by NIH grant RFA-PH-05-126. This research (for the first author) 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 authors and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE. Permission to use the data was obtained through data use agreement #189 from TRICARE Management Activity.

host of difficulties in achieving such performance (as described by Fienberg and Shmueli (2005)), foremost among them the seasonal, nonstationary, and autocorrelated nature of the health data being monitored. Although current biosurveillance data are typically monitored at a daily frequency, the methods and results in this paper are general and apply to data at other time scales as well.

In order to identify outbreaks in pre-diagnostic health data, most modern algorithms use some type of forecasting and then monitor the residuals (i.e., forecast errors) using a control chart. When a sequence of residuals suggests a disease outbreak, an alert is generated. In several cases this sequence of forecast, then monitor is not done explicitly. Instead, a control chart is altered and then applied to the raw data. Even in such cases, the algorithm can be represented as a combination of forecasting and control chart monitoring. For example, in EARS or BioSense (programs initiated by the Centers for Disease Control and Prevention), a control chart is applied to the raw data, but a 'sliding window' of recent data is used to set the control limits (as suggested in Hutwagner, Thompson, Seeman and Treadwell (2003)). This combination is equivalent to using a moving-average forecast method to forecast the next point and then applying a simple control chart to the forecast errors. ESSENCE (a Department of Defense monitoring system) uses regression to forecast the next day's value, and then explicitly monitors the residuals in a control chart (described in Lombardo, Burkom and Pavlin (2004)).

The effect of forecasting precision on detection rate is therefore applicable to biosurveillance, since it is important to know how much benefit improved forecasting will provide. Forecast methods have several properties which are useful in biosurveillance aside from their precision, such as robustness to non-

normality, outliers, or outbreaks in the training data, as well as generating uncorrelated residuals (discussed later in Section 4.4.1). When faced with a new forecast method which is more precise but is worse in these properties, the improvement must be quantified to understand the tradeoff. Although central to many applications, the effect of forecast precision on detection performance has not been directly studied. Monitoring and forecasting have been discussed as being similar in purpose and approach (by Atienza, Ang and Tang (1997)). The two also have been used together for the opposite purpos:; rather than using forecasting to improve control chart detection, control charts have been used to identify issues in the forecast method, starting with Van Dobben De Bruyn (1967). In the following sections, we examine the quantitative effect of forecasting improvement on control chart detection, both in the standard case of independent normal residuals as well as under various violations of assumptions which occur in practice.

2 Problem description

Our ultimate purpose is to provide early notice of an outbreak based on finding an outbreak signature in the data. We will often refer to the outbreak signature as simply the 'outbreak'. However, it should be clear that there is a distinction between the outbreak itself and its manifestation or signature in the monitored data series. For evaluation purposes, algorithms must be evaluated on their ability to detect these outbreak signatures. In the following we first describe the metrics used to evaluate the performance of a biosurveillance algorithm. We then describe control charts, which are the most popular tool used for monitoring forecast errors.

2.1 Performance metrics

Consider a time series of health data, collected periodically. Daily is the most common collection interval, and we use the convention of assuming daily collection throughout the paper; however, the results apply equally well for different intervals. Now consider that we have many such series of the same type; some contain outbreaks, and some do not. Consider that we take k series, each with an outbreak, and m series without an outbreak. The main metrics used in biosurveillance to evaluate an outbreak detection method are sensitivity, specificity, and timeliness. However, we use the less ambiguous measures described in Fricker, Jr., Hegler and Dunfee (ming), which are closer to those used in classical SPC:

Detection Rate : the proportion of outbreaks detected, out of the k series with outbreaks. As k is made arbitrarily large, this measures the per-outbreak probability that there will be an alert sometime during the outbreak.

ATFS : the Average Time to False Signal, this is the average number of days until an alert, over the m series without outbreaks. As m is made arbitrarily large, this measures the expected time until a false alert. For implementations which reset after any alert, 1/ATFS will be the average proportion of days with false alerts, given that there is no outbreak.

ATFOS : the Average Time to First Outbreak Signal, this is the expected number of days until an alert is generated, given that the method does eventually alert during the outbreak signal.

We note here that each of these metrics depends on the outbreak signal itself, as well as on the underlying health data series. In biosurveillance the variety of data sources leads to a variety of baseline behaviors. Furthermore, the exact outbreak signature is unknown. Therefore, it is generally important to consider a variety of baseline time series as well as a variety of outbreak signal shapes and sizes for evaluating algorithm performance. Given the wide array of possibilities, simulation methods, and metrics, it is difficult to make overall claims about the performance of one method versus another. However, if we can generate general claims about the effect of forecast precision on detection effectiveness, it will allow us to rank methods based on their actual forecast effectiveness, independent of the outbreak type or monitoring method. More importantly, quantifying this effect allows us to determine *how much* more effective the better forecast method will be, specific to the type of monitoring being applied and the type and size of the outbreak to be detected. In addition, by examining properties of the residuals, we can identify those cases where a better forecast method will *not* necessarily produce better detection.

2.2 Overview of control charts

Control charts are statistical tools for monitoring process parameters and alerting when there is an indication that those parameters have changed. Originally designed for use in manufacturing, they are now widely used in health-related fields, particularly in biosurveillance (as seen in Benneyan (1998); Woodall (2006)). There are some difficulties in directly applying control charts to daily pre-diagnostic data, since classical control charts assume that observations are independent, identically distributed, and typically normally dis-

tributed (or with a known parametric distribution). For this reason, forecasting should be used to precondition the data in order to create residuals which better meet control chart requirements.

Control charts are usually two-sided, monitoring for an increase or decrease in the parameter of interest. This is done using an upper control limit (UCL) and lower control limit (LCL), respectively. In biosurveillance, we are usually only concerned with a significant *increase* in the underlying behavior indicative of a disease outbreak, and therefore only a UCL is used. The control chart is applied to a sample statistic (often the individual daily count), and alerts when that statistic exceeds the UCL. This UCL is a constant, set to achieve a certain false alert level; the true alert rate can then be computed.

The three main types of control charts are the Shewhart, Cumulative Sum (CuSum), and Exponentially Weighted Moving Average (EWMA). These are covered in detail in Montgomery (2001). One point to remember is that in biosurveillance, the CuSum and EWMA are reset after an alert. This is done because the false alert rate determines the amount of resources which must be devoted to a system. Resetting ensures that the ATFS is both the average time to first false signal and the average time between false signals; thus the overall false alert rate will be 1/ATFS, even though the rate will not be constant for each day.

3 Problem formalization

We first consider a series with no outbreak signals; we call such a series the *underlying background* or *baseline* data, denoted as u_t (t=1,2,...). It is this

underlying background that a forecast method is attempting to forecast. The predictions from the forecast method are f_t; if we examine the forecast errors, $e_t = y_t - f_t$, we can estimate the Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) and bias of those errors. This will be useful in evaluating detection effectiveness.

Since we do not actually know a priori whether or not the data contain an outbreak, we denote the actual values in the series as y_t. When there is no outbreak signal, $y_t = u_t$. Let o_t be the outbreak signal at time t. In general, $y_t = u_t + o_t$, which assumes an additive number of cases due to the outbreak signal. For most days, $o_t = 0$, whereas $o_t > 0$ only on days where there is an outbreak. This reflects the epidemiological model commonly used in biosurveillance. If a multiplicative outbreak effect is assumed ($y_t = u_t \cdot o_t$, where $o_t > 1$ only on days where there is an outbreak), we can model $log(y_t)$ instead of y_t, thereby converting to an additive outbreak form.

Since we do not know if an outbreak is present in a given series, we will refer to the difference $r_t = y_t - f_t$ simply as a residual, rather than a pure forecast error. In the absence of an outbreak signal, r_t will be a pure forecast error and the residuals will have variance equal to the forecast method's MSE (assuming unbiased forecasts). However, in the presence of an outbreak signal, r_t will contain an additional term; since the forecast method is forecasting only the underlying background, we will not call this a forecast error. The residual can thus be separated into two components, $r_t = (u_t - f_t) + o_t$. The first component is the forecast error ($e_t = u_t - f_t$) and the second is the outbreak signal (o_t).

An illustration of these components can be seen in Figure 1. It shows the original series and forecasts in the left panel; the residuals obtained from

subtracting the forecasts from the original series are shown in the right panel.

(Figure 1 approximately here.)

4 Theoretical performance

4.1 Standard Gaussian, known variance, one-day 'spike' outbreak signal

In our analysis, we first assume that the forecast method generates forecast errors with a given MSE. Initially, we assume that these errors are independent, normally distributed, with mean 0 and constant variance. We later relax these assumptions and re-evaluate performance.

We now consider an additive outbreak signal that is injected into the monitored series. This outbreak signal is considered to be independent of the background or residuals. Thus, we are in the realm of standard control charts: we are seeking a change in the process mean, given a series of independent identically distributed (iid) normal observations. Let us first consider a single-day 'spike' outbreak signal.

Note that when converting a time series to a series of residuals, if the residuals have 0 mean, then the residuals' variance is equal to the forecast method's MSE.

First, consider a Shewhart chart being applied to residuals that are iid, $e_t \sim N(0, \sigma^2)$. Setting the control limit at UCL means that a false alert will occur

with ATFS/

$$ATFS = \frac{1}{1 - \Phi(UCL/\sigma)}. \quad (1)$$

In the simplest case, the outbreak signal is of constant size, $o_t = \eta$. In this case, the algorithm will detect if $e_t/\sigma + \eta/\sigma > UCL/\sigma$. By using the same transformation as above, the control chart will correctly alert on the day of the outbreak if $Z > UCL/\sigma - \eta/\sigma, Z \sim N(0,1)$, which translates into a probability of detection equal to

$$DetectionRate = 1 - \Phi(UCL/\sigma - \eta/\sigma). \quad (2)$$

Note that we obtain Equation (1) by setting $\eta = 0$.

Now consider two forecast methods, f_1 and f_2, with RMSEs equal to σ_1 and σ_2, respectively, and where $\sigma_1 < \sigma_2$ (i.e., forecast method f_1 provides more precise forecasts). If detectors on each of f_1 and f_2 are set to have the same false alert rate ($ATFS_1 = ATFS_2$) we can write $UCL_1/\sigma_1 = UCL_2/\sigma_2 = a$. Since $\sigma_1 < \sigma_2$, then clearly $UCL_1 > UCL_2$. Thus the corresponding probabilities of detection will be $TA_1 = 1 - \Phi(a - \eta/\sigma_1)$ and $TA_2 = 1 - \Phi(a - \eta/\sigma_2)$. Because $\sigma_1 < \sigma_2$, we get $TA_1 > TA_2$ and therefore the more precise forecast method (f_1) will also provide a higher Detection Rate.

The effects are shown in Figure 2, where the Detection Rate of five forecast methods are compared, all normalized to have the same ATFS. We see that as the forecasting becomes more precise (i.e., the RMSE decreases), the detection rate increases. While this relationship is monotonic (a lower RMSE always results in improved detection), the amount of improvement depends on the *size* of the outbreak signal (η). Since $UCL = \sigma\Phi^{-1}(1-1/ATFS)$ (see equation

1), the improvement in detection rate from using f_1 over f_2 can be expressed as

$$\Phi(\Phi^{-1}(1-1/ATFS)-\eta/\sigma_2)-\Phi(\Phi^{-1}(1-1/ATFS)-\eta/\sigma_1). \tag{3}$$

Due to the nature of the normal cumulative distribution function Φ, this quantity must be computed numerically.

(Figure 2 approximately here.)

For an EWMA chart, we compute similar probabilities in Appendix A.

4.2 Stochastic outbreak signal

Thus far, the assumptions are still in the realm of standard control charts. A slightly more general case is to assume that the outbreak is not of fixed size, but is instead stochastic, e.g., $o_t \sim N(\eta, \nu^2)$. In this case, a Shewhart chart has probability of detection equal to

$$DetectionRate = 1-\Phi\left((UCL-\eta)/\sqrt{\sigma^2+\nu^2}\right). \tag{4}$$

Figure 3 shows the relationship between expected outbreak size (η) and Detection Rate for a stochastic outbreak signal, applying a Shewhart control chart to five forecast methods with varying RMSEs. It can be clearly seen that, compared to the fixed-size spike, the increased variance in the outbreak signal reduces the detection rate for larger spikes, but increases it for smaller ones; this effect is proportional to the amount of outbreak-size variance, ν^2. In comparing two methods, this distortion can drastically affect the relative performance of the two forecast methods: *a large advantage of one forecast*

method over another under one variance may be almost trivial under a different outbreak-size variance. However, this significant effect due to the stochastic nature of an outbreak is seldom if ever considered.

(Figure 3 approximately here.)

4.3 Multi-day outbreaks

When outbreak signals last more than one day, there are more chances to detect them. This allows consideration not only of the probability of detection, but also the distribution of *when* the outbreak is detected.

We first consider a fixed step increase of size η that starts at time i and continues indefinitely ($o_i = \eta, \forall i > t$). Such an outbreak signal could be the result of an environmental contamination (biological or chemical) resulting in a constant increase in the number of illness cases. Since any control chart method will eventually alert, we focus on timeliness over true alert probabilities. In control chart terminology, this is usually referred to as the Average Run Length (ARL), which is the expected number of days until an alert is generated.

For the Shewhart chart, each day is essentially a Bernoulli trial in terms of detection, with probability of success $p = 1 - \Phi(UCL/\sigma - \eta/\sigma)$. Thus, the number of days until detection is a geometric random variable with expected value $ARL = (1-p)/p$. (If the alerting day is considered to be included, this is $(1-p)/p + 1 = 1/p$.)

The relationships between outbreak size and expected delay (i.e., the number of days until detection), for forecast methods of varying precision, can be

seen in Figure 4. Results for CuSum and EWMA charts, which are better suited for detecting small step increases, can be seen in Appendix B. Note that the *quantity* of the performance difference varies significantly based on the outbreak size and the amount of forecast improvement; the quantity is crucial in determining the benefits from using an improved forecast method.

(Figure 4 approximately here.)

We caution that in practice the expected value (ATFOS) may not be the most useful metric, since it will incorporate alerts generated many days after the outbreak signal first appeared in the data. In other words, it averages over the entire distribution of possible delays. If a detection must occur within the first k days of an outbreak signal to be useful to the user, then more effective metrics of model performance and comparison would be the probability of alert *within the first k days* and the *conditional expected timeliness*, given that an alert occurred within the first k days. This same issue comes up when recognizing the finite duration of outbreaks; if an outbreak only lasts k days, then a detection must certainly occur within k days to be useful. In essence, one must make sure to examine detection probability as the probability of practically useful detection, and timeliness as the expectation of delay, conditional on a practically useful detection.

An important condition of our results regarding improved forecasting leading to improved detection is that the forecast method does not include the outbreak in the background data and effectively forecast the combination (a problem described in Burkom, Murphy and Shmueli (2007)). This can be achieved in practice by using a 'guardband window' which means that forecasts are generated for more than one day ahead. Forecasting farther into the future

generally results in reduced precision, which in turn leads to deteriorated detection probabilities and timeliness. It is, in fact, precisely when considering tradeoffs of this kind that one must know the quantitative loss from decreased forecast precision.

4.4 Performance under assumption violation

In practice, it is rare that forecast methods will provide residuals that are iid normal with mean 0. There are two major types of violations that appear in residuals from biosurveillance data: autocorrelation and seasonal (cyclical) variance (e.g., Lotze, Murphy and Shmueli (2008); Burkom, Murphy and Shmueli (2007)). We now examine the relationship between detection and forecast precision under the two types of assumption violations.

4.4.1 Autocorrelation

Autocorrelation in a series of residuals means that the residuals on consecutive days are not independent. Autocorrelated residuals indicate that the forecast method did not capture part of the dependence structure in the raw data (such as a seasonal component). In biosurveillance data, the most pronounced autocorrelation in series of residuals is that of lag 1 (the correlation between r_t and r_{t-1}) and it is typically positive. When we refer to autocorrelation hereafter, we are referring to positive autocorrelation.

When data are autocorrelated, the series will have increased variance due to the autocorrelation. In the case of an autoregressive model of order 1 (AR(1)),

given by

$$y_t = \phi y_{t-1} + \epsilon_t, \epsilon_t \ N(0, \sigma_z^2) \tag{5}$$

the resulting variance is $\sigma_z^2/(1-\phi^2)$ (from Maragah and Woodall (1992)). The effect of autocorrelation on detection performance has been examined in the control chart literature. Several papers that look at Shewhart, CuSum, and EWMA charts applied to autocorrelated series indicate that autocorrelation leads to a greater number of false alarms, due to the greater variance in the series (described in Maragah and Woodall (1992); Woodall and Faltin (1993); Padgett, Thombs and Padgett (1992); Noorossana and Vagjefi (2005)) . However, for Shewhart charts, if the control chart limits are adjusted to account for the variance of the actual autocorrelated series (rather than the variance which would exist without any autocorrelation), then the overall probability of detection will remain the same for a spike outbreak. We do caution that while this is true unconditionally, the probabilities of detection, conditional on the value for the previous day, are not identical for each day. The probability of alert will be larger on days following large values, and smaller on days following small values. As we discuss below, this implies that methods taking this conditional probability into account should provide improved detection performance.

Although the performance of a Shewhart chart is unaffected by autocorrelation on single-day spike outbreaks, there will be a longer average delay in detection when considering a multi-day outbreak signal, both for Shewhart and other control charts. When the oubtreak begins on a day with a small residual, which is too low to alert (even after the outbreak signal addition), the subsequent residuals will likely also be too low for the outbreak to be detected. Thus,

average delay will increase for higher values of autocorrelation. These effects are shown in Section 5.1.

To determine whether or not a residual series contains autocorrelation, an autocorrelation plot may be used (with α-level bounds $(z_{1-\alpha/2})/\sqrt{N}$). When autocorrelation is present, as mentioned above, the conditional probability of detection varies by day; this implies that one might use an ARMA-type or other model as an additional forecasting step on the residuals from the original forecast method (such mdoels are described in Box and Luceno (1997); Montgomery and Mastrangelo (1991)). However, note that in the case of multi-day outbreaks, such models will incorporate the outbreak signal into the forecasting, and thus the assumption of independence of outbreak and forecast error will be violated. The results of such incorporation on the performance of detection algorithms is discussed in Hong and Hardin (2005). The decrease in performance from incorporating an outbreak must be measured against the gain achieved by reducing the autocorrelation, as mentioned at the end of Section 4.3; it is precisely these kinds of tradeoffs for which this theoretical quantification is useful.

4.4.2 Seasonal variance

When the forecast precision is non-constant, even if the forecast method produces unbiased forecasts, the theoretical analysis in Section 4.1 does not hold. This can occur, for example, when the series of daily counts follows a Poisson distribution with different λ parameters for each day of the week. A similar effect can occur when an additive forecast method is applied to a series with multiplicative background behavior. Although a preliminary log trans-

formation of the series may be a reasonable and popular solution, such a transformation will also have a significant impact on the outbreak signal.

Seasonal variance can also be induced by deseasonalizing methods which normalize values by multiplication. An example is deseasonalizing a series from a day-of-week effect using the ratio-to-moving-average method (as described in Lotze, Murphy and Shmueli (2008)). Conversely, if such methods are used appropriately, they may help reduce seasonal variance by normalizing the variance of residuals across seasons. However, here too there is the danger that a transformation that affects the variance of the residuals will also impact the size of the outbreak signal.

If there is periodic variance in the residuals series with period k, we can represent the variance as a set of variances, $\sigma_1^2 ... \sigma_k^2$. Then the overall variance of the series (assuming that the mean residual=0 for each season) is $\sum_{i=1}^{k} (1/k)\sigma_i^2$. If the seasonal pattern is such that some days have equal variance, we can represent this as $\sum_{i=1}^{k} \alpha_i \sigma_i^2$, where α_i is the proportion of days with variance σ_i^2. Given this mixture model for seasonal variance, we can compute the probability of detection. For a step outbreak signal using a Shewhart control chart, we can compute separate probabilities of detection by season; thus, the probability of detection for an outbreak signal of size η is

$$DetectionRate = \sum_{i=1}^{k} \alpha_i P(detection|\eta, \sigma_i). \tag{6}$$

Using Equation (2), this quantity is equal to $\sum_{i=1}^{k} \alpha_i (1 - \Phi((UCL/\sigma_i) - (\eta/\sigma_i)))$, where the UCL is derived from the overall variance of the series.

Consider, for example, a series with seasonal variance between weekdays and weekends, such that weekday residuals have a higher variance than that on

weekends. The detection probability is therefore $(5/7)(1-\Phi(UCL/\sigma_{weekday}-\eta/\sigma_{weekday}))+(2/7)(1-\Phi(UCL/\sigma_{weekend}-\eta/\sigma_{weekend}))$. If the overall variance is kept constant at 100, but the difference between weekend and weekday variance is increased, the performance becomes more markedly different from the constant variance case. We can see this difference in performance in Figure 5; as weekday and weekend variances become more distinct, detection rates deteriorate for small outbreak sizes, but actually *improve* for some intermediate outbreak sizes. At these intermediate outbreak sizes, the increased probability of detection when the outbreak occurs on low-variance weekends outweighs the decrease in performance on higher-variance weekdays. As the overall variance is increased, this "kink" pattern of deviation from the constant variance case is increased.

(Figure 5 approximately here.)

In conclusion, if variance is strongly differentiated by season, an improved RMSE will not always give better detection performance, depending on the size of the outbreak. For some outbreak sizes, a forecast method with a larger overall RMSE but low weekend RMSE can outperform a forecast method with a smaller overall RMSE. When there is significant seasonal variance, the performance can be evaluated more accurately using Equation 6 and estimates for the different seasonal variances. This suggests that improved monitoring can be achieved by using different UCLs and/or different forecast methods for each season.

5 Empirical validation

We have shown theoretical results for detection and timeliness under different forecast methods and outbreaks. Next we describe simulation experiments that demonstrate the effects of autocorrelation. We then evaluate the applicability of the theoretical results on authentic biosurveillance data.

5.1 Autocorrelation simulation

To study the impact of autocorrelation on detection and timeliness performance, residuals were simulated using different levels of autocorrelation, but again maintaining the same overall series variance. In the Shewhart charts using spike outbreaks, no significant deviation was seen from the theoretical performance, when the control limit was set according to the final resulting variance. Figure 6 shows that the detection performance is not affected by autocorrelation. Figure 7 shows a significant deterioration in timeliness for small outbreak sizes and high autocorrelation. This is in agreement with Wheeler (1991, 1992) regarding the relatively small impact of most autocorrelation levels on Shewhart chart performance.

(Figure 6 approximately here.)

(Figure 7 approximately here.)

5.2 Authentic biosurveillance data with simulated outbreaks

An authentic health dataset is now used to determine the effectiveness of theory when estimating performance of currently-used forecast methods. These tests show the applicability of the theory to the evaluation of forecast methods on actual health data for detecting siaw'aw outbreaks. If the predicted performance and actual performance match well, then the theoretical analysis can be used to accurately estimate the detection performance of actual systems; thus, the forecast metrics can be a useful comparison metric, without requiring computationally intensive simulation studies.

To examine the forecast methods' effectiveness, authentic health series data are used, with a simulated outbreak signal inserted at various possible dates of outbreak. This methodology is now commonly used in biosurveillance to estimate the effectiveness of detection (Goldenberg, Shmueli, Caruana and Fienberg (2002); Reis and Mandl (2003); Stoto, Fricker, Jain, Davies-Cole, Glymph, Kidane, Lum, Jones, Dehan and Yuan (2006); Burkom, Murphy and Shmueli (2007)). See Shmueli and Burkom (ming) for more examples. The technique involves using an authentic health data set from a health provider, simulating a potential outbreak signal and inserting the simulated additional counts in the authentic data. Then, the detection algorithm is run to determine whether it alerts during the simulated outbreak, and if so, how quickly. By repeating this routine multiple times and inserting the simulated outbreak at multiple points, one can estimate how the detection algorithm would perform during an actual outbreak.

Our authentic dataset comes from the BioALIRT program conducted by the

U.S. Defense Advanced Research Projects Agency (DARPA), described in Siegrist and Pavlin (2004). It includes three types of daily counts: 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 10 U.S. metropolitan areas with substantial representation of each data type. For this study, we use the daily count of respiratory symptoms from civilian physician office visits, all within a particular U.S. city. The data consist of counts from 700 days, from July 1, 2001 to May 31, 2003, and can be seen in Figure 8. The first 1/3 of the data (233 days) was used for training, and the last 2/3 (467 days) for testing.

(Figure 8 approximately here.)

Simulated spike outbreak signals of various sizes (0-300) were generated and inserted into every day in the test set, creating 467 trials for each outbreak signal size. For each outbreak size, the detection rate was calculated as the average over all 467 insertions. An illustration of the process can be seen in Figure 9.

(Figure 9 approximately here.)

Three forecast methods for forecasting next-day daily counts were compared:

1. Holt-Winter's multiplicative exponential smoothing, with a seasonal component which captures day-of-week (7 seasons) and smoothing parameters equal to alpha=0.4, beta=0, gamma=0.15, plus a restriction to not update parameters when a day's percentage error is greater than 50%. Burkom, Murphy and Shmueli (2007) have shown that this method is effective in the context

of biosurveillance, as well as being easy to understand and apply for a large class of data types. Little data history is needed, and due to its highly adaptive nature, it reduces the need for individual modifications for specific data sources and syndrome groupings.

2. A regression model, which models the log of daily counts as a linear combination of three types of predictor terms: a linear trend, day-of-week indicators, and sin and cos terms with a yearly period for capturing yearly seasonality. Each day's prediction is given by

$$f_t = exp\left(\hat{\beta}_{0t} + \hat{\beta}_{1t}t + \hat{\beta}_{dt} + \hat{\beta}_{st}sin(t * \frac{2\pi}{365.25}) + \hat{\beta}_{ct}cos(t * \frac{2\pi}{365.25})\right) \quad (7)$$

Where $\hat{\beta}_{dt}$ is the day-of-week coefficient for day t, and all $\hat{\beta}$ values are estimated by minimizing least squares over days 1...t-1.

3. 7-day differencing, as proposed in Muscatello (2004), which models the next day's count as the count from one week previous.

For a more detailed description of these methods and comparison of their performance, see Lotze, Murphy and Shmueli (2008). For each method, the first 1/3 of the data (233 days) was used for training, and the last 2/3 (467 days) for comparison. Note, however, that the 7-day differencing has no real "training" to speak of, and that both the Regression and Holt-Winters method incorporate *all* previous days when making a forecast. A small plot of the residuals from each forecasting method can be seen in Figure 14.

The RMSE for each forecast method was computed and used to generate a theoretical performance curve for each forecast method as in Section 4.1. Actual performance was computed using the method described in Figure 1,

using the forecast method for prospective forecasting, subtracting the forecast to generate residuals, and applying a Shewhart control chart to those residuals.

(Figure 10 approximately here.)

Results can be seen in Figure 10, which compares the actual performance from a forecasting method's residuals to the performance which would be expected from the theoretical performance for residuals of the same overall RMSE. When an overall UCL was used, the actual performance was somewhat similar to that predicted by theory, but seemed to underdetect small outbreaks and overdetect midsized outbreaks. This result is similar to that seen under seasonal variance (see Section 4.4.2), which reflects the seasonal variance of the residuals (seen in Figure 11).

(Figure 11 approximately here.)

A further examination was done, with variance computed for each day-of-week and performance predicted using seasonal variance computations. The results are shown in Figure 12, where an improved fit is seen, especially for the Holt-Winters residuals, although there is still some difference on the larger outbreaks.

(Figure 12 approximately here.)

In order to compare timeliness, the experiment was repeated using step outbreaks instead of spike outbreaks. Step outbreaks have an additional report count which begins on a certain day, and lasts indefinitely. Figure 13 compares the timeliness performance of real forecast methods to theoretical performance predicted by a 7-day seasonal variance model. The timeliness is worse for small outbreaks, particularly for the regression and 7-day differencing.

(Figure 13 approximately here.)

The extra delay for regression and 7-day differencing seems to be due to autocorrelation: as seen in Figure 14, the regression and 7-day differencing residuals have larger autocorrelation than Holt-Winters. Alternatively, the overall differences may be due to the bias of the residuals (none has mean 0) or their non-normal distribution. However, we see that the forecast methods' performance ranking is related to their RMSE ranking, as expected.

(Figure 14 approximately here.)

In short, the effect of forecast precision on detection performance for these health data is close to that expected; more precise forecast methods result in improved detection, accounting for seasonal variance improves performance estimation, and the amount of difference between forecast methods depends on outbreak size.

6 Conclusions and future work

In this paper, we have shown that improved forecasting results in improved detection, both in terms of probabilities of true alert and in timeliness. We examined the effect of forecast precision on detection performance theoretically and quantified the effects under standard control chart assumptions. We have also examined the effects of assumption violation on this relationship, showing that improved forecasting does not always result in improved detection, as in the case of seasonal variance. We conclude that forecasting should be tuned to best capture the background non-outbreak behavior, while detection should be tuned to the outbreak signal. However, the level of investment in

more precise forecasts should be weighed against factors such as the required outbreak size, amount of residual autocorrelation, and risks of the forecast method capturing the outbreak.

Several questions arise for practical consideration. First, while we have explored the effects of autocorrelation and seasonal variance, we have not explored the effects of biased or non-normal residuals. As we have seen in the authentic data, biases can arise in actual residuals and can affect performance. In addition, while we have examined the detection performance for spike outbreaks and timeliness performance for step outbreaks, a complete delay distribution would include both metrics and give a more complete picture; it would also be relevant to consider average and complete delay distributions for other outbreak shapes, such as exponential or lognormal rise. Lastly, we have not considered the quality of the training data used for prediction. Not only should it be possible to apply previous work to give expected performance based on the amount of training data (such as the multiplicative Holt-Winters accuracy bound given by Chatfield and Yar (1991)), but the impact of outbreaks contaminating the training data or different guardband widths should also be considered.

In conclusion, given the forecasting precision needed for useful detection, the question is whether that level of precision is achievable. This raises the question of whether there is enough quality of signal in pre-diagnostic data. The random elements in the data impose a limit on how well we can forecast, how low an RMSE we can achieve, and ultimately on how well we can detect. It may be that, due to the high noise in most pre-diagnostic data, relatively high false alert rates are required in order to detect outbreaks in a timely manner. For example, if the desired performance is to have a false alert once every

two weeks, and have a 95% chance of detecting a spike outbreak impacting 100 people, to achieve this one would need normal residuals with a forecast $RMSE < 32$. In contrast, the best forecast method used here has $RMSE = 59$ on actual data. If we cannot accept a higher false alert rate, then we must either find a way to further improve our forecast methods (e.g., by incorporating other sources of information or by using ensembles, as in Lotze and Shmueli (2008)), or tailor our detectors to specific outbreak signals.

7 Figures

References

Atienza, O., Ang, B., and Tang, L. (1997). Statistical process control and forecasting. *International Journal of Quality Science* **2,** 37–51.

Benneyan, J. C. (1998). Statistical quality control methods in infection control and hospital epidemiology, part ii: Chart use, statistical properties and research issues. *Infection Control and Hospital Epidemiology* **19(4),** 265–283.

Box, G. and Luceno, A. (1997). *Statistical Control: By Monitoring and Feedback Adjustment.* Wiley-Interscience, 1st edition.

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007). Automated time series forecasting for biosurveillance. *Statistics in Medicine* **26,** 4202–4218.

Chatfield, C. and Yar, M. (1991). Prediction intervals for multiplicative holt-winters. *Internaional Journal of Forecasting* **7,** 31–37.

Crowder, S. V. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics* **29,** 401–407.

Fienberg, S. E. and Shmueli, G. (2005). Statistical issues and challenges as-

sociated with rapid detection of bio-terrorist attacks. *Statistics in Medicine* **24(4),** 513–529.

Fricker, Jr, R. D., Hegler, B. L., and Dunfee, D. A. (2008). Comparing syndromic surveillance detection methods: Ears versus a cusum-based methodology. *Statistics in Medicine* **27(17),** 3407–29.

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* **99,** 5237–5240.

Hong, B. and Hardin, J. M. (2005). A study of the performance of multivariate forecast-based surveillance schemes for infectious diseases on multiple locations. In *Proceedings of the Joint Statistical Meeting, Minneapolis, Minnesota*.

Hutwagner, L., Thompson, W., Seeman, G., and Treadwell, T. (2003). The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health* **80 (2) Suppl,** 89–96.

Lombardo, J. S., Burkom, H., and Pavlin, J. (2004). Essence ii and the framework for evaluating syndromic surveillance systems. *MMWR* **53(Suppl),** 159–165.

Lotze, T., Murphy, S. P., and Shmueli, G. (2008). Preparing biosurveillance data for classic monitoring. *Advances in Disease Surveillance* .

Lotze, T. H. and Shmueli, G. (2008). Ensemble forecasting for disease outbreak detection. In *Proceedings of the 23rd AAAI Conference on Artficial Intelligence (AAAI-08)*.

Maragah, H. D. and Woodall, W. H. (1992). The effect of autocorrelation on the retrospective x-chart. *Journal of Statistical Computation and Simulation* **40,** 29–42.

Montgomery, D. C. (2001). *Introduction to Statistical Quality Control*. John Wiley & Sons, third edition.

Montgomery, D. C. and Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology* **23,** 179–204.

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

Noorossana, R. and Vagjefi, S. J. M. (2005). Effect of autocorrelation on performance of the mcusum control chart. *Quality and Reliability Engineering International* **22(2),** 191–197.

Padgett, C. S., Thombs, L. A., and Padgett, W. J. (1992). On the -risks for shewhart control charts. *Communications in StatisticsSimulation and Computation* **21,** 1125–1147.

Reis, B. and Mandl, K. (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making* **3,**.

Shmueli, G. and Burkom, H. S. (forthcoming, expected 2008). Statistical challenges in modern biosurveillance. *Technometrics (Special Issue on Anomaly Detection)* .

Shu, L., Jiang, W., and Wu, S. (2007). A one-sided ewma control chart for monitoring process means. *Communications in Statistics - Simulation and Computation* **36:4,** 901–920.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag.

Siegrist, D. and Pavlin, J. (2004). Bio-alirt biosurveillance detection algorithm evaluation. *Morbidity and Mortality Weekly Report (MMWR)* **53,** 152–158.

Stoto, M., Fricker, R. D., Jain, A., Davies-Cole, J. O., Glymph, C., Kidane,

G., Lum, G., Jones, L., Dehan, K., and Yuan, C. (2006). Evaluating statistical methods for syndromic surveillance. In Wilson, A., Wilson, G., and Olwell, D. H., editors, *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance and Biometric Authentication*, pages 141–172. ASA-SIAM.

Van Dobben De Bruyn, C. S. (1967). The interplay of tracking signals and adaptive predictors. *The Statistician* **17(3),** 237–246.

Wheeler, D. (1991). Shewhart's charts: Myths, facts and competitors. In *ASQC Quality Congress Transactions*. Milwaukee, WI.

Wheeler, D. J. (1992). Correlated data and control charts. In *Fifth Annual Forum of the British Deming Association*.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology* **38(2),** 89–104.

Woodall, W. H. and Faltin, F. W. (1993). Autocorrelated data and spc. *ASQC Statistics Division Newsletter* **13,** 18–21.

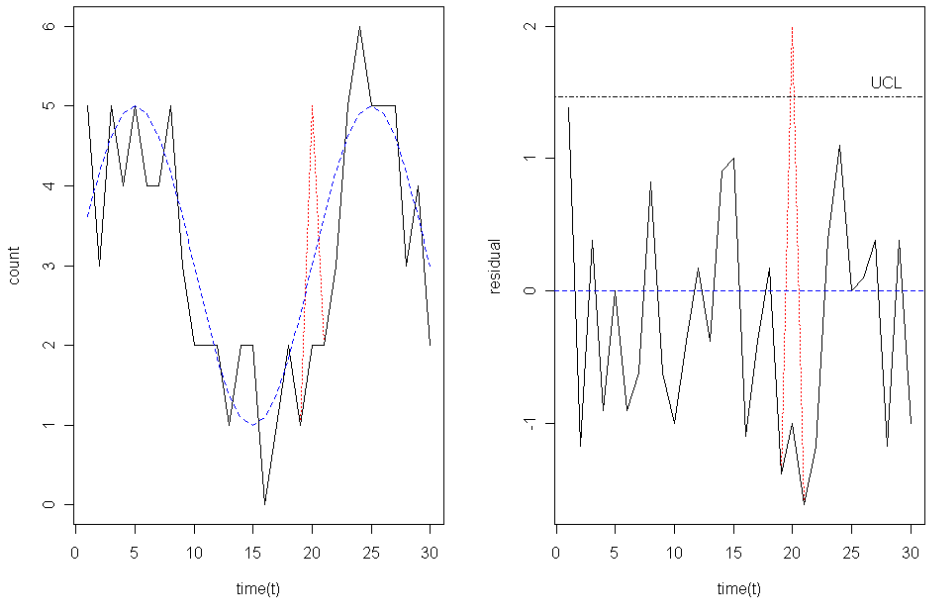

Fig. 1. An original series (solid line, u_t) and its forecasts (dashed line, f_t) are shown in the left panel; the residuals from subtracting forecasts from the series are shown in the right panel, in a one-sided Shewhart control chart. The dotted line is the addition of an outbreak signal ($o_t + u_t$).

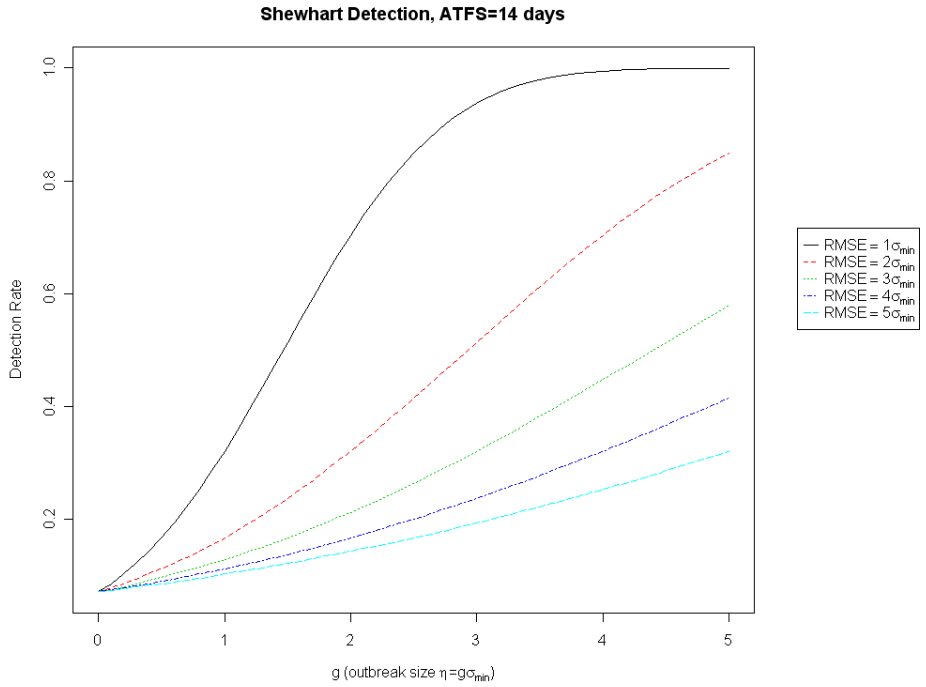

Fig. 2. Comparing Shewhart chart performance for forecast methods with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method)

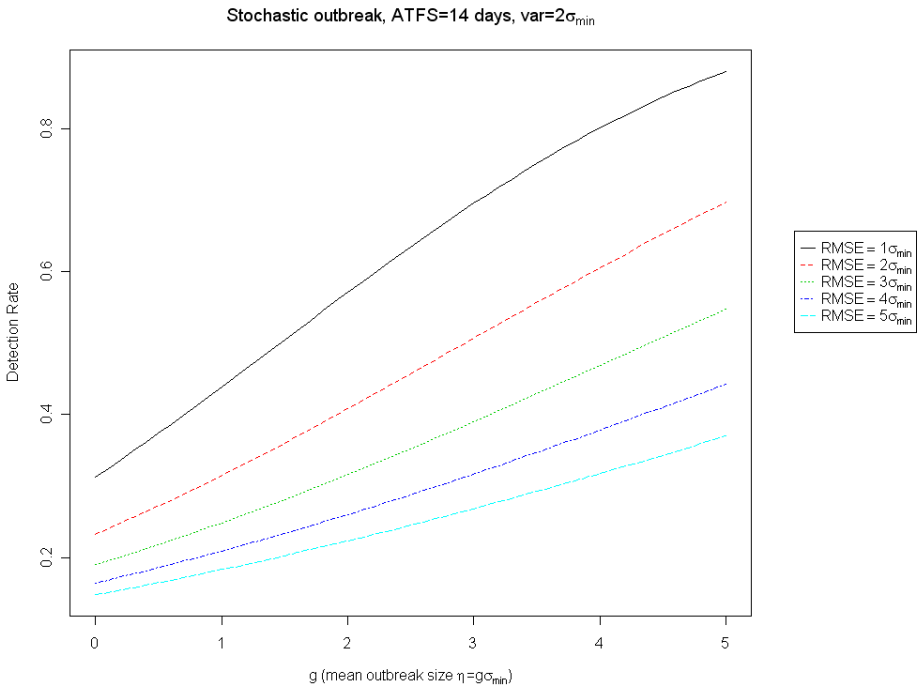

Fig. 3. Comparing Shewhart chart performance for forecast methods with different RMSEs, for stochastic outbreak, as a function of outbreak mean size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method). Note that intercepts at $g = 0$ are not equal, due to the stochastic nature of the outbreak.

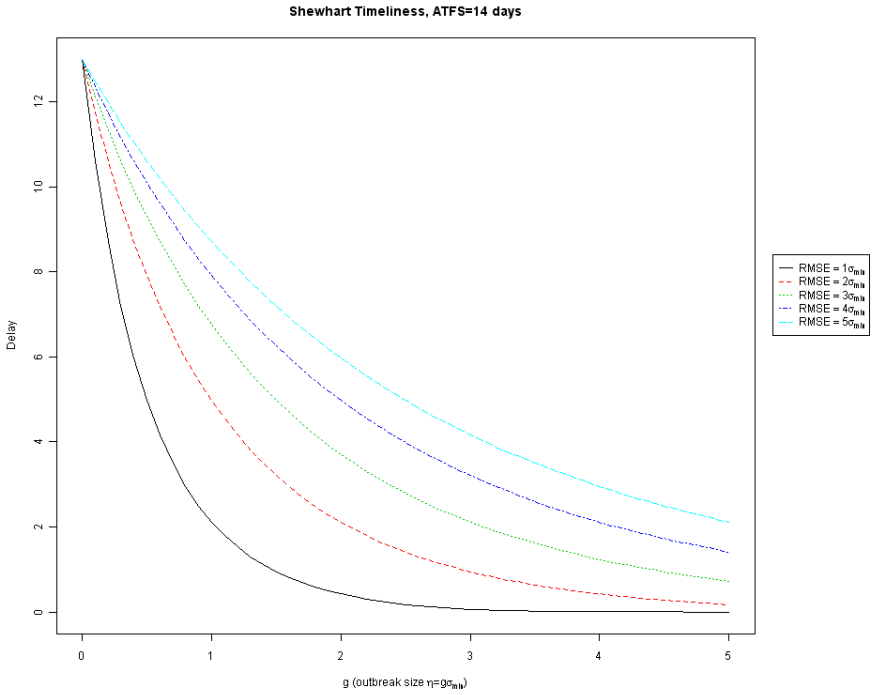

Fig. 4. Comparing Shewhart chart timeliness for forecast methods with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method)

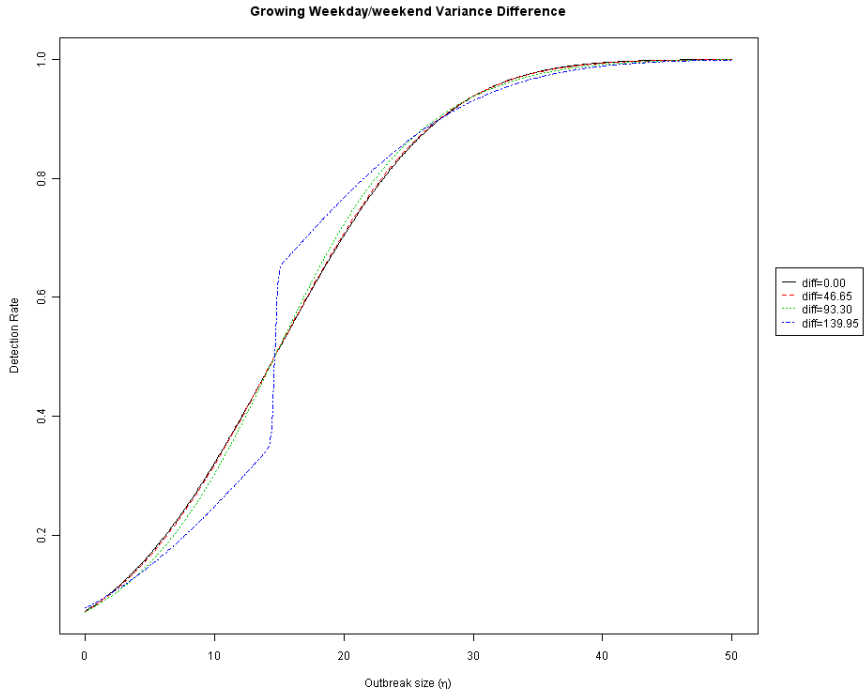

Fig. 5. Comparing Shewhart chart performance for forecast methods with different residual seasonal variances (diff=difference between weekday and weekend residual variance) but identical overall variance $\sigma^2 = 100$

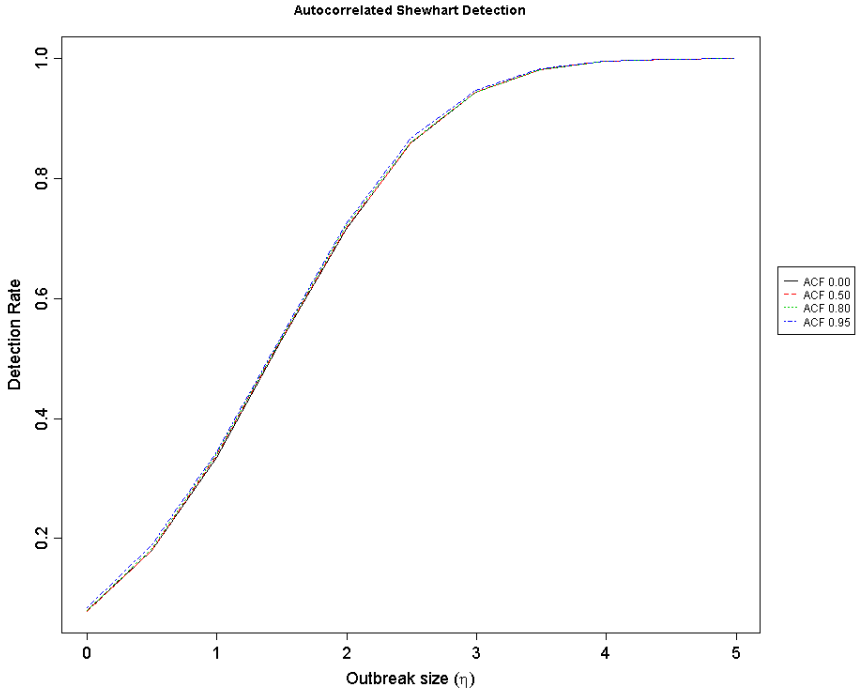

Fig. 6. Comparing Shewhart chart performance for forecast methods with different residual autocorrelation levels (ACF) but identical overall variance $\sigma^2 = 1$)

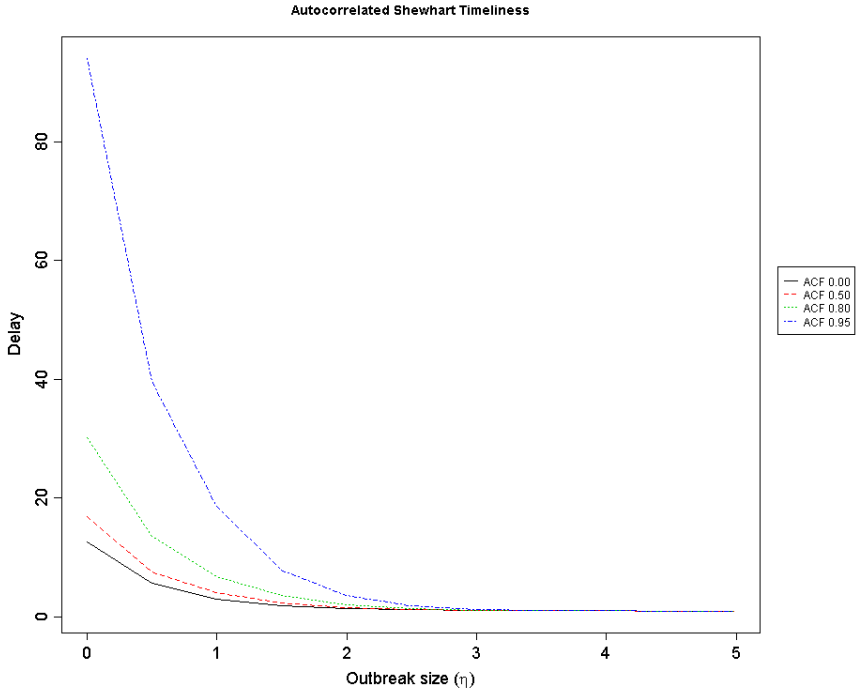

Fig. 7. Comparing Shewhart chart timeliness for forecast methods with different residual autocorrelation levels (ACF) but identical overall variance $\sigma^2 = 1$)

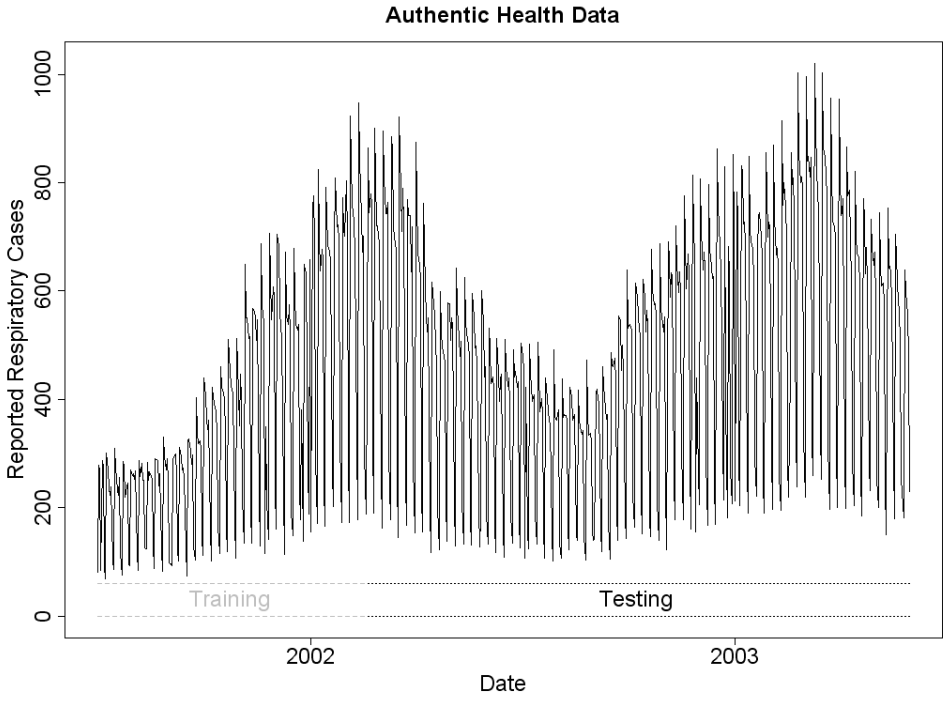

Fig. 8. The original, raw respiratory health data series, with training and testing sections labeled.

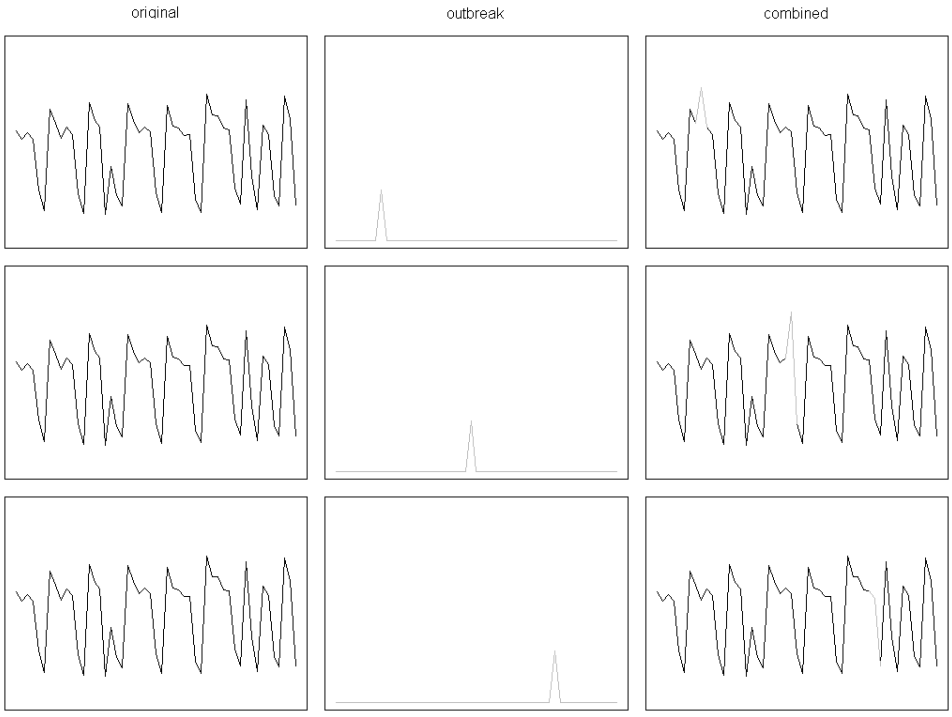

Fig. 9. An illustration of taking raw, authentic health data series and injecting a spike outbreak into three different days, resulting in three test data series. These test series are then used as outbreak-labeled time series for estimating the method's detection rate. In our implementation, 467 such data series were created for each outbreak size.

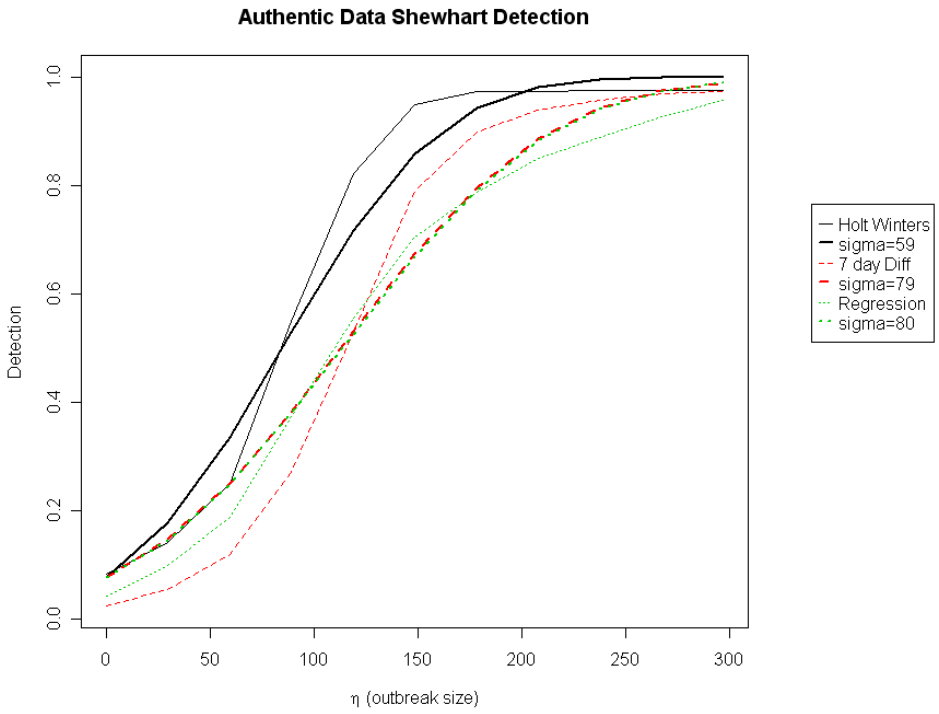

Fig. 10. Comparing actual (thin) and theoretical (thick) Shewhart chart performance for forecast methods with different RMSEs, assuming constant variance, as a function of outbreak size (η). Solid=Holt-Winters, Dashed=7-day Diff, Dotted=Regression. Each forecasting method has the sigma for its residuals measured, and is matched with a plot of theoretical performance for residuals of the same sigma.

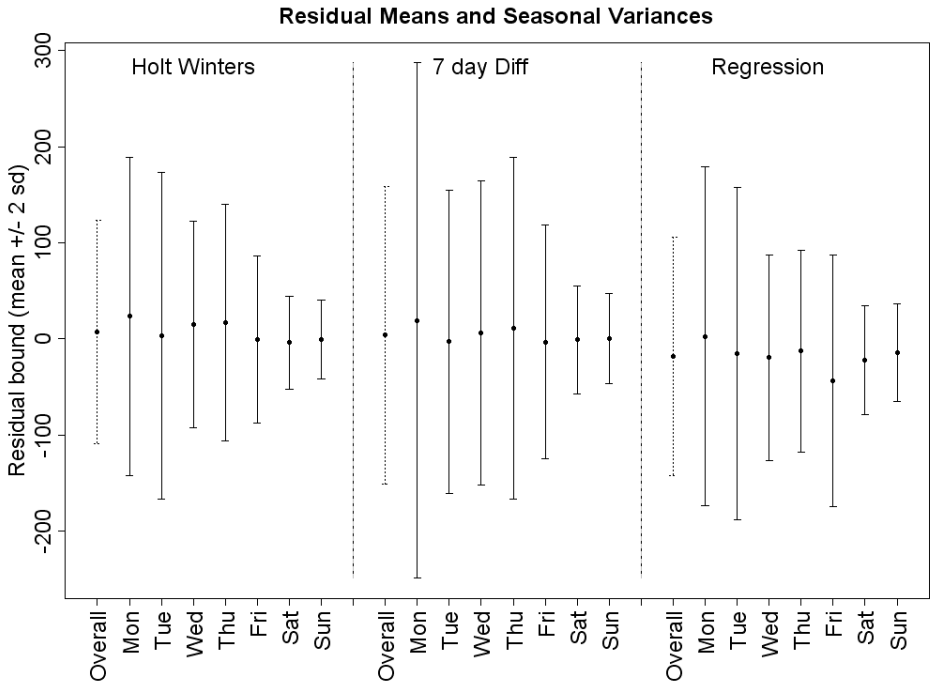

Fig. 11. Overall residual variance of the three forecast methods, and variance by day-of-week. Seasonal day-of-week variance affects detection performance, and can be accounted for using the formulas in Section 4.4.2.

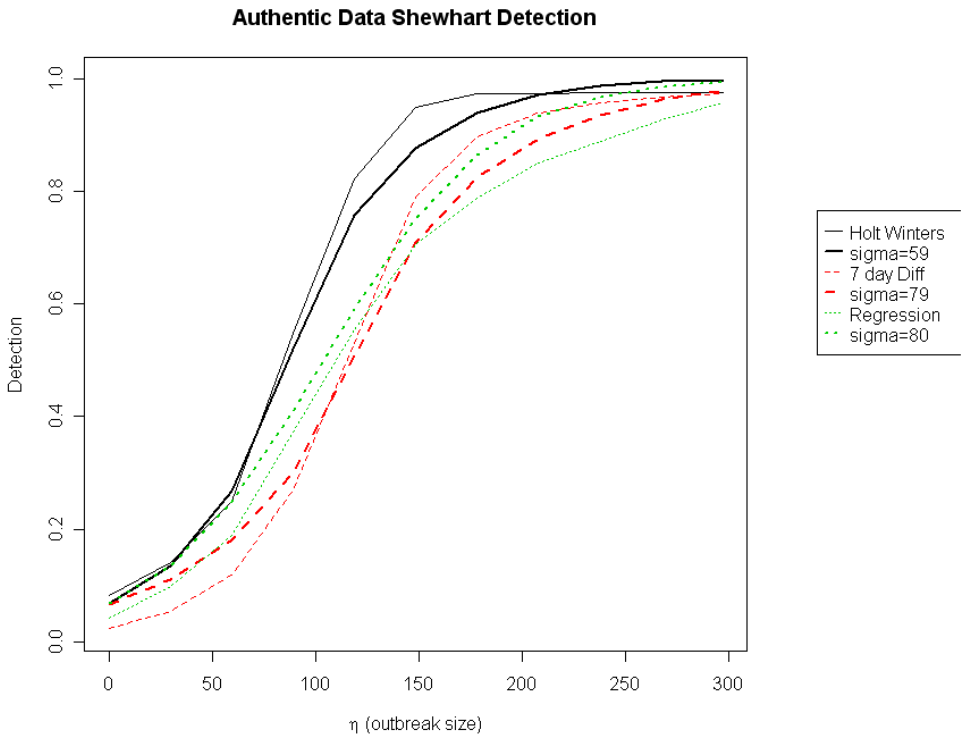

Fig. 12. Comparing actual (thin) and theoretical (thick) Shewhart chart performance for forecast methods with different RMSEs, assuming day-of-week variance, as a function of outbreak size (η). Solid=Holt-Winters, Dashed=7-day Diff, Dotted=Regression. Each forecasting method has the sigma for its residuals measured, and is matched with a plot of theoretical performance for residuals of the same sigma, with the same day-of-week residual variance.

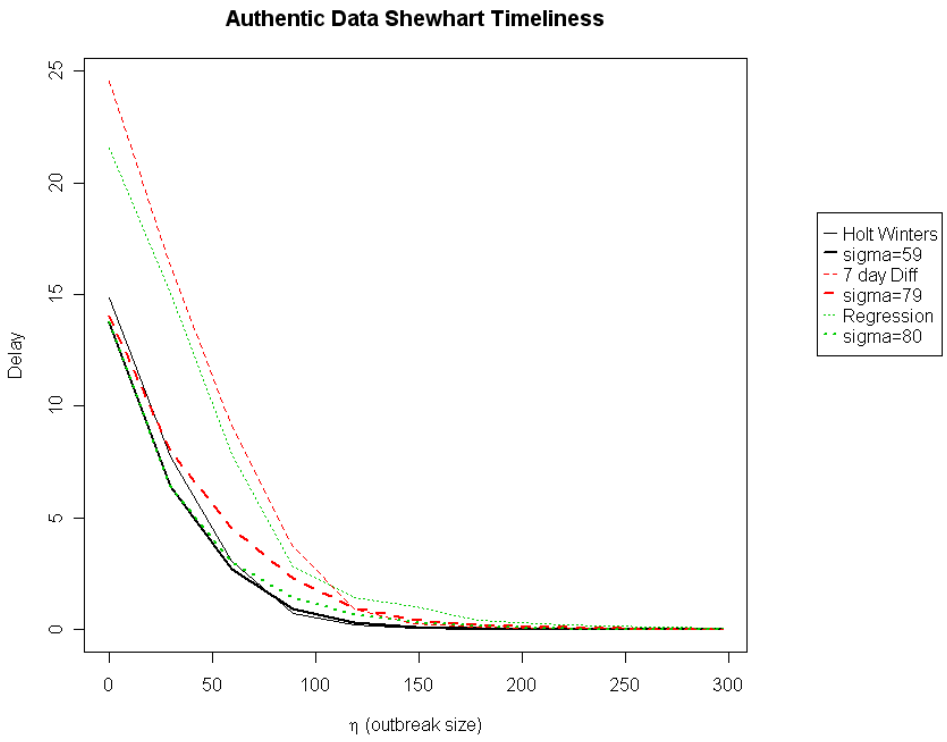

Fig. 13. Comparing actual (thin) and theoretical (thick) Shewhart chart timeliness for forecast methods with different RMSEs, assuming constant variance, as a function of outbreak size (η). Solid=Holt-Winters, Dashed=7-day Diff, Dotted=Regression. Each forecasting method has the sigma for its residuals measured, and is matched with a plot of theoretical performance for residuals of that sigma.

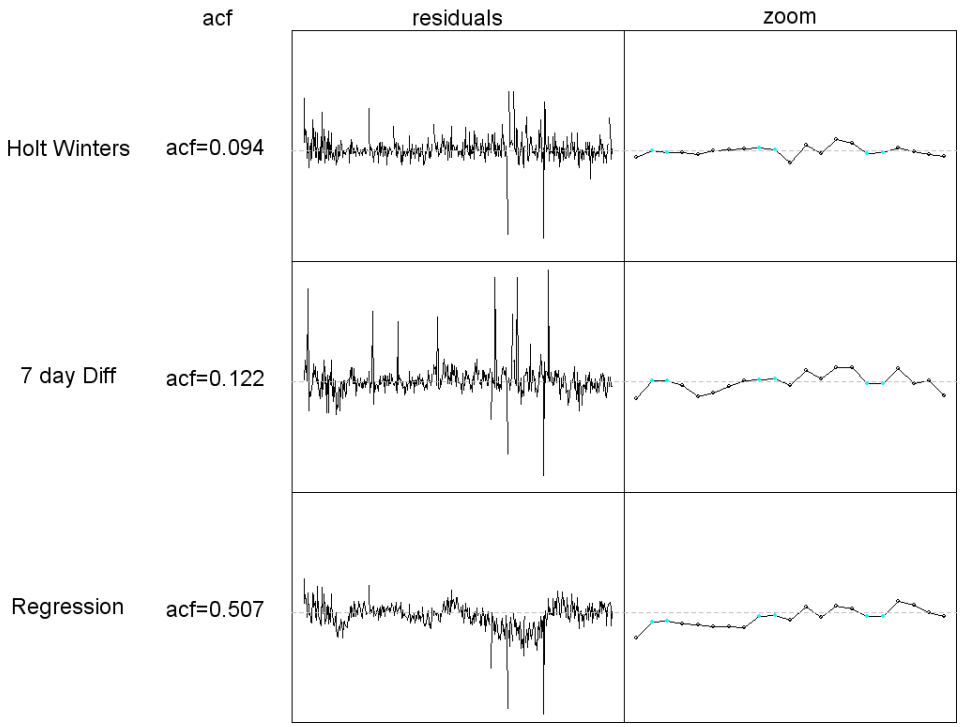

Fig. 14. Examining the residual autocorrelation of the three forecast methods. Y-axes are the same for all graphs, which show the overall residuals for each forecasting method and a zoomed-in portion to show daily detail.

A Appendix A: EWMA performance

We can measure the effect of improved forecasting on EWMA detection, as in Equation (2) for Shewhart charts, by noting that $EWMA_t$ is a normal random variable, with mean 0 and variance as in Montgomery (2001):

$$\sigma^2_{EWMA_t} = \sigma^2 \left(\frac{\lambda}{2-\lambda} \right) \left(1 - (1-\lambda)^{2t} \right), \tag{A.1}$$

where t is the number of time points since the EWMA was started. After an initial startup period, the variance converges to $\sigma^2 (\lambda/(2-\lambda))$. The one-sided EWMA has been shown to have very similar performance to the EWMA approximated by this steady-state normal distribution (Shu, Jiang and Wu (2007)). By similar argument to the Shewhart case (in section 4.1), we can show that the improvement in detection probability from using f_1 over f_2 can be expressed as

$$\Phi \left(\Phi^{-1}(1 - \frac{1}{ATFS}) - \frac{\lambda\eta}{\sigma_2 \sqrt{\frac{\lambda}{2-\lambda}}} \right) - \Phi \left(\Phi^{-1}(1 - \frac{1}{ATFS}) - \frac{\lambda\eta}{\sigma_1 \sqrt{\frac{\lambda}{2-\lambda}}} \right) \tag{A.2}$$

Figure A.1 shows the relationship between outbreak size (η) and Detection Rate for EWMA detectors when applied to five different forecast methods, each with a different RMSE.

(Figure A.1 approximately here.)

Comparing Figures 2 and A.1 shows that an EWMA chart has a lower chance of detecting a spike outbreak compared to a Shewhart chart with the same ATFS, when both are applied to residuals with the same RMSE. The reason is that by giving the maximum weight to the most recent observation, the

Shewhart chart is more tuned to detect spike outbreak signals. A much larger spike is necessary to achieve the same detection rate with an EWMA chart. However, we also note that as a weighted sum of observations, the EWMA chart is more robust to deviations from normality, and so may be more effective when the residual distribution is further from normal.

B CuSum and EWMA timeliness

In a CuSum chart, the monitoring statistics on different days are no longer independent, and therefore the number of days until an alert is no longer a geometric variable. However, the ATFS can still be accurately determined using numerical methods or approximations. One such approximation is found in Siegmund (1985), which approximates the ATFS by:

$$ATFS \approx 2(e^{-2(UCL/\sigma+1.166)} + UCL/\sigma + .166) \tag{B.1}$$

This same approximation can provide the ATFOS:

$$ATFOS \approx \frac{e^{-2\Delta b} + 2\Delta b - 1}{2\Delta^2}, \tag{B.2}$$

where $\Delta = \eta/\sigma - 1/2$ and $b = UCL/\sigma + 1.166$.

For the EWMA chart, the ARL is computed numerically; we use the method described in Crowder (1987), numerically integrating the Fredholm equation using Gaussian quadrature.

The relationships between outbreak size and expected delay (i.e., the number of days until detection), for forecast methods of varying precision, can be seen in Figures B.1 and B.2, for CuSum and EWMA charts respectively.

(Figure B.1 approximately here.)

(Figure B.2 approximately here.)

For each of these methods, more precise forecasts result in faster detection. One surprising result, as seen in Figure B.3, is that although the CuSum chart has improved detection over the Shewhart chart for small outbreak signals (as expected), the Shewhart chart quickly catches up and outperforms the CuSum as the outbreak size increases. In addition, this timeliness improvement appears to be bounded below, and to hold only for a certain range of outbreak sizes.

(Figure B.3 approximately here.)

B.1 Autocorrelation

Results using CuSum charts on autocorrelated data are similar to those for Shewhart charts (Section 5.1). Detection may be slightly affected for spike outbreaks, and timeliness is more strongly affected than in Shewhart charts.

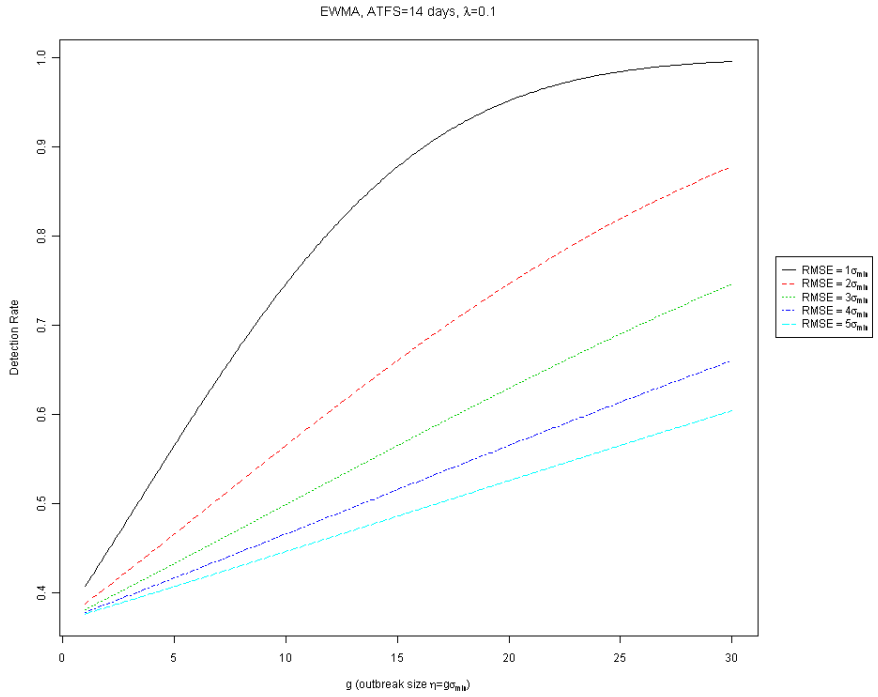

Fig. A.1. Comparing EWMA chart performance for forecast methods with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method.)

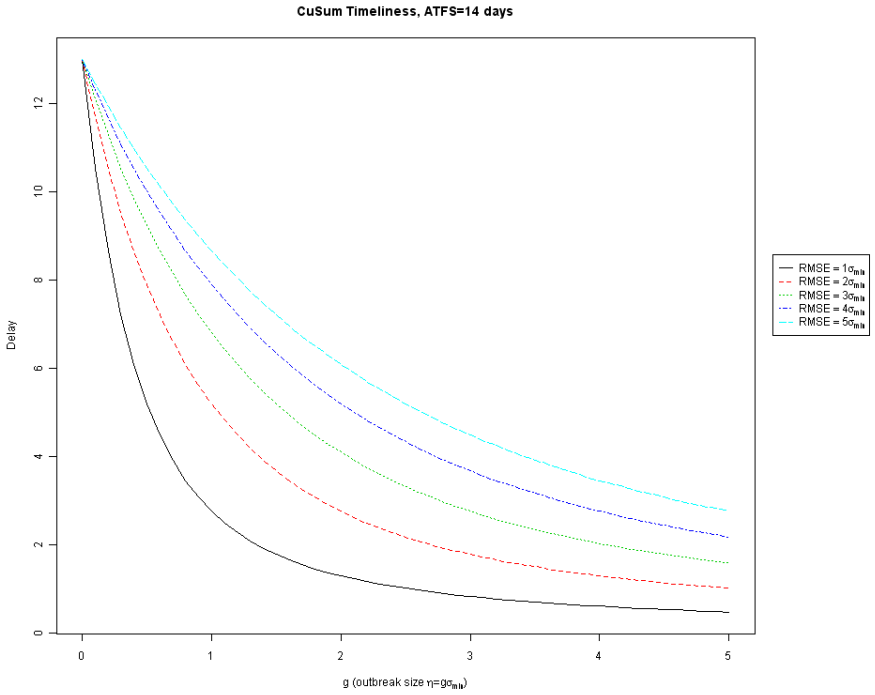

Fig. B.1. Comparing CuSum chart timeliness for forecast methods with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method.)

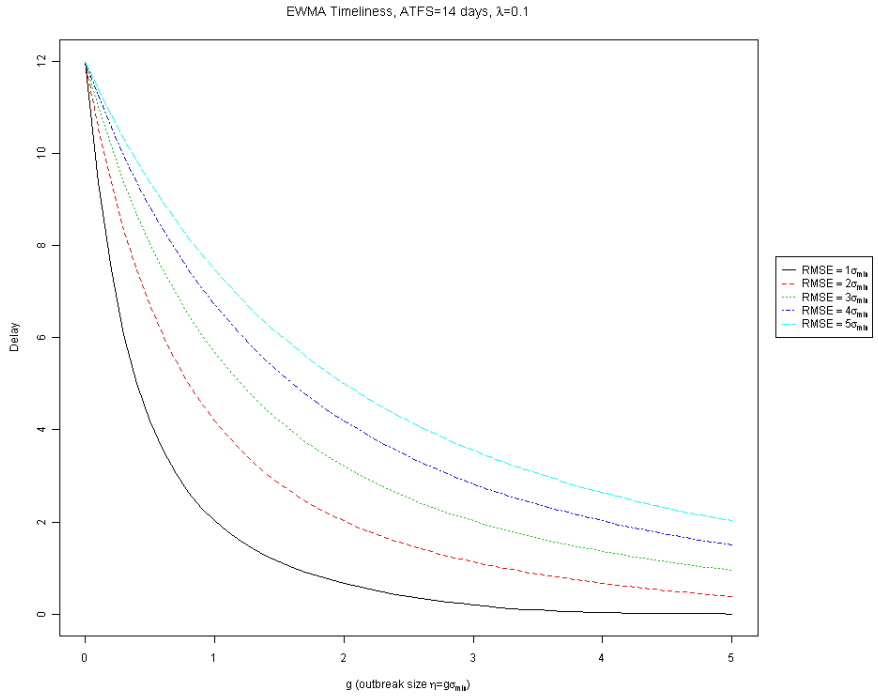

Fig. B.2. Comparing EWMA chart timeliness for forecast methods with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecast method.)

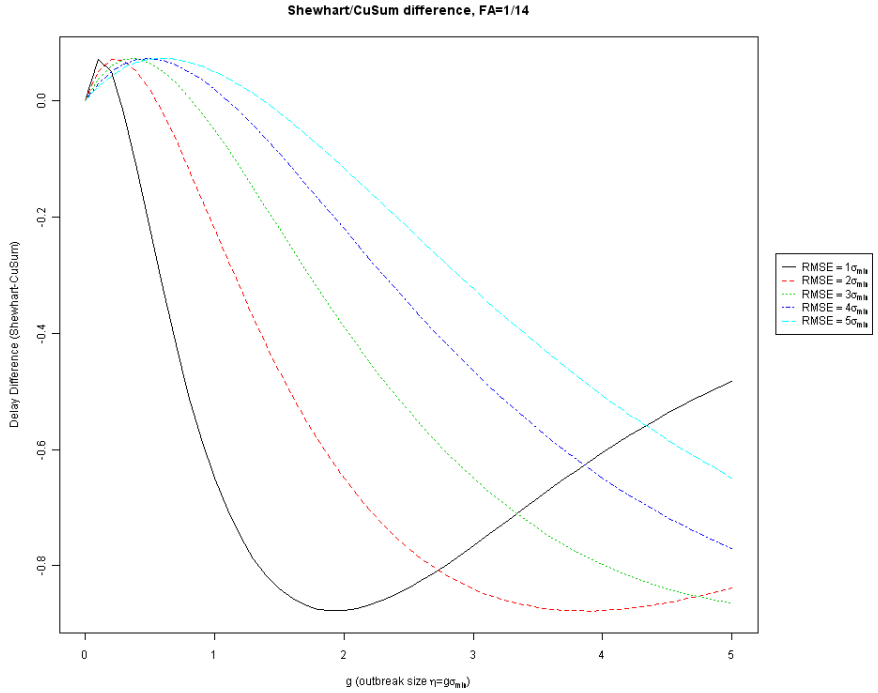

Fig. B.3. Expected difference in delay resulting from using a Shewhart instead of a CuSum, on the same forecast residuals.