On the relationship between forecast accuracy and detection performance: An application to biosurveillance

Thomas Lotze
Applied Mathematics and Scientific Computation Program
University of Maryland
College Park, MD 20742
Email: lotze@math.umd.edu

Galit Shmueli
Dept. of Decision, Operations & Information Technologies
and Center for Health and Information Decision Systems
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. In this paper, we describe the relationship between forecast accuracy and the detection accuracy of a method.**

We argue that most existing temporal detection methods for biosurveillance can be characterized as a forecasting component coupled with a monitoring/detection component. We show that improved forecasting results in improved detection and we quantify the relationship between forecast accuracy and detection metrics under different scenarios. The forecast accuracy can then be used to rate an algorithm's expected performance in detecting outbreaks. Simulation is used to compare empirical performance with theoretical results; we also show examples with authentic biosurveillance data.

***Index Terms*—Anomaly detection, Biosurveillance, Control Charts**

I. 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 alarm rate. There are a host of difficulties in achieving such performance (1), foremost among them the seasonal, non-stationary, and autocorrelated nature of the health data being monitored. Because of this, most modern algorithms use some type of forecasting and then monitor the residuals (i.e., forecast errors) using a control chart. 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.

II. 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.

A. Performance Metrics

The main metrics used in biosurveillance to evaluate an outbreak detection method are:

- True alarm rate (TA): probability of alert, given that there is an actual outbreak signal (per outbreak)
- False alarm rate (FA): probability of alert when no outbreak signal is present (per day)
- Timeliness: expected number of days until an alert is generated, given that an alert is generated

Our claims will therefore be along these lines: given certain conditions on the forecast errors, certain conditions on the outbreak signal, and a certain false alert rate, an algorithm will have a minimum probability of detection (or timeliness). This probability of detection will depend on the accuracy of the forecaster, and will improve when an improved forecaster is used.

B. 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 (2; 3). 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 distributed (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. In biosurveillance, we are usually only concerned with a significant *increase* in the underlying behavior, and therefore only an upper control limit (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. Control chart limits, as used in this context, are thus a way of employing prediction bounds as repeated statistical tests of parameter shift.

In this paper, we consider the one-sided Shewhart chart. This chart monitors the series Y_t ($t = 1, 2, \ldots$) itself, alerting when $Y_t > UCL_{Shewhart}$, where Y_t is the observation (e.g., the daily count) at time t. $UCL_{Shewhart}$ is often set to $\mu + 3\sigma$, where σ is the standard deviation of the series. Under an assumption of a normal distribution, this scheme results in a false alert rate of $FA = P(Y_t > \mu + 3\sigma) = 0.0013$. The one-sided Shewhart chart is most effective at detecting a medium-to-large single-point spike increase in the mean. In the journal version of this paper, CuSum and EWMA charts are also considered, as well as bounding methods when the residual distribution is unknown.

III. Problem Formalization

We first consider a series with no outbreak signals; we will call such a series the underlying background or baseline data, denoted as u_t (t=1,2,...). It is this underlying background which a forecaster is attempting to forecast. The predictions from the forecaster are p_t; if we examine the forecast errors, $e_t = u_t - p_t$, 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.

However, since we do not actually know a priori whether or not the data contain an outbreak, we take the actual values in the series to be 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.

Since we do not know if an outbreak is present in a given series, we will refer to the difference $r_t = y_t - p_t$ 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 forecaster's MSE (assuming unbiased forecasts). However, in the presence of an outbreak signal, the residual can thus be separated into two components, $r_t = (u_t - p_t) + o_t$. The first component is the forecast error ($e_t = u_t - p_t$) and the second is the outbreak signal (o_t). An illustration of these components can be seen in Figure 1.

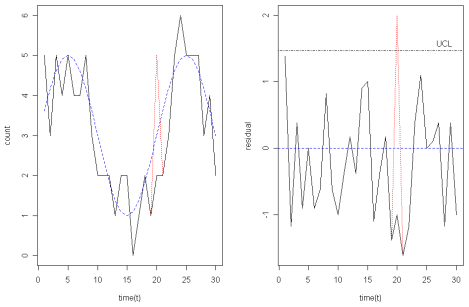

Fig. 1. An original series (solid line, u_t) and its forecasts (dashed line, p_t) are shown in the left panel; the residuals from subtracting forecasts from the series are shown in the right panel, in a one-sided Shewhart control chart. The dotted spike is the addition of an outbreak signal ($o_t + u_t$).

IV. Theoretical Performance

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

In our analysis, we first assume that the forecaster 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. Let us first consider a single-day 'spike' outbreak signal, which the standard Shewhart chart is most effective at detecting.

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 forecaster's MSE.

In this case, $e_t \sim N(0, \sigma^2)$. Setting the control limit at UCL means that a false alarm will occur if $e_t/\sigma > UCL/\sigma$. Since $Z = e_t/\sigma \sim N(0, 1)$, a false alarm will occur if $Z > UCL/\sigma$, which translates into a probability of false alert equal to

$$FA = 1 - \Phi(UCL/\sigma). \quad (1)$$

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 alarm if $Z > UCL/\sigma - \eta/\sigma$, which translates into a probability of detection equal to

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

By varying the UCL, both the FA and TA rates are altered, thereby creating a Receiver Operating Characteristic (ROC) curve.

Now consider two forecasters, f_1 and f_2, with RMSEs equal to σ_1 and σ_2, respectively, and where $\sigma_1 < \sigma_2$ (i.e., forecaster

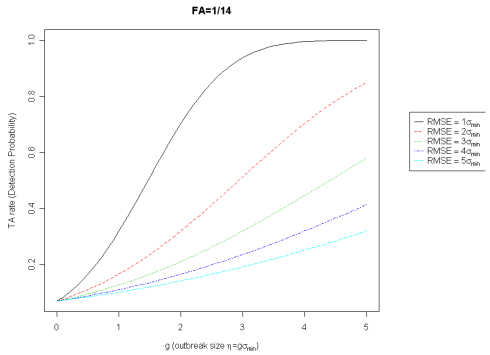

Fig. 2. Comparing Shewhart chart performance for forecasters with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecaster.) Smaller RMSEs result in improved detection.

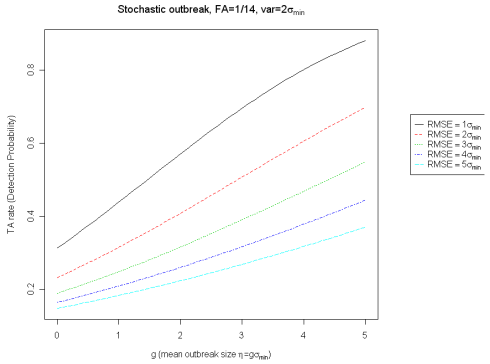

Fig. 3. Comparing Shewhart chart performance for forecasters with different RMSEs, for stochastic outbreak, as a function of outbreak mean size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecaster.)

f_1 provides more precise forecasts). If detectors on each of f_1 and f_2 are set to have the same false alarm rate ($FA_1 = FA_2$) 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 forecaster (f_1) will also provide improved detection (i.e., higher TA).

The effects are shown in Figure 2, where the TA of five forecasters are compared, where all forecasters have the same FA. We see that as the forecasting becomes more precise (i.e., the RMSE decreases), detection probability 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 $UCL = \sigma\Phi^{-1}(1-FA)$ (see equation 1), the improvement in detection probability from using f_1 over f_2 can be expressed as

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

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

B. 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., normally distributed according to some mean and variance. We still assume that it is independent from the underlying background. More formally, we assume $o_t \sim N(\eta, \nu^2)$. In this case, $r_t = o_t + e_t \sim N(\eta, \nu^2 + \sigma^2)$, or equivalently, $Z = (r_t - \eta)/\sqrt{\sigma^2 + \nu^2}) \sim N(0, 1)$. A Shewhart control chart will correctly alert if $Z > (UCL - \eta)/\sqrt{\sigma^2 + \nu^2}$, which translates into a probability of detection equal to

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

The false alarm rate remains the same, since the background behavior has not changed.

We can construct plots of the relationship between alerting and outbreak size for a stochastic outbreak similar to Figure 2. Figure 3 shows the relationship between expected outbreak size (η) and TA for a stochastic outbreak signal, applying a Shewhart control chart to five forecasters with varying RMSEs.

In Figure 3 we see that, compared to the fixed-size spike, the increased variance in the outbreak signal reduces the detection probability for larger spikes, but increases it for smaller ones; this effect is proportional to the amount of outbreak-size variance, ν^2.

The distortion due to the stochastic nature of the outbreak signal significantly reweights the detection performance on different outbreak sizes. In comparing two methods, this distortion can drastically affect the relative performance of the two forecasters. A large advantage of one forecaster over another under a certain variance may be almost trivial under a different outbreak-size variance.

C. Multi-day Outbreaks

When outbreak signals last more than one day, there are more chances to detect them. This makes it possible to consider not only the probability of detection, but also the distribution of *when* the outbreak is detected (the distribution of timeliness).

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.

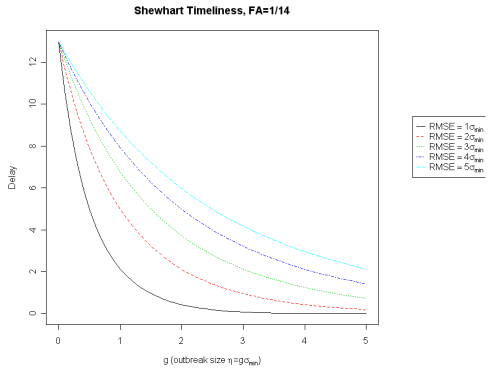

Fig. 4. Comparing Shewhart chart timeliness for forecasters with different RMSEs, as a function of outbreak size (g=η/σ_{min}, where σ_{min} is the RMSE of the best forecaster.)

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 $ARL = (1-p)/p$.

For each of these methods, more precise forecasts result in faster detection. We caution, however, that in practice the expected value (ARL) may not be the most useful metric, since it incorporates alerts generated many days after the outbreak signal first appears 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. 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.

D. Performance Under Assumption Violation

In practice, it is rare that forecasters will provide residuals that are iid normal with mean 0. There are two main types of violations one would expect to see in biosurveillance data: autocorrelation and seasonal (cyclical) variance (e.g., (4; 5)). We now examine the relationship between detection and forecast precision under the two types of assumption violations.

1) Autocorrelation: Autocorrelation in a series of residuals means that the residuals on consecutive days are not independent. Autocorrelated residuals indicate that the forecaster 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 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

$$y_t = \phi y_{t-1} + \epsilon_t, \tag{5}$$

the resulting variance is $\frac{\sigma_z^2}{1-\phi^2}$ (6). The control chart literature has examined the effect of autocorrelation on detection performance. Several papers (6; 7; 8) indicate that autocorrelation leads to a greater number of false alarms, due to the greater variance in the series. But 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 probability of detection will remain the same for a spike outbreak. When an outbreak signal lasts longer than one day, however, there will be a longer average delay in detection for Shewhart and other control charts.

2) Seasonal Variance: When the forecast precision is non-constant, even if the forecaster produces unbiased forecasts, the theoretical analysis above will 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. 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 (4).

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 $\sum_{i=1}^{k} \frac{1}{k}\sigma_i^2$. 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

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

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

Consider for example a series with seasonal variance between weekdays and weekends. 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; performance is worsened for small outbreak sizes, but actually *improved* for some intermediate outbreak sizes. As the overall variance is increased, this "kink" pattern of deviation from the constant variance case is increased.

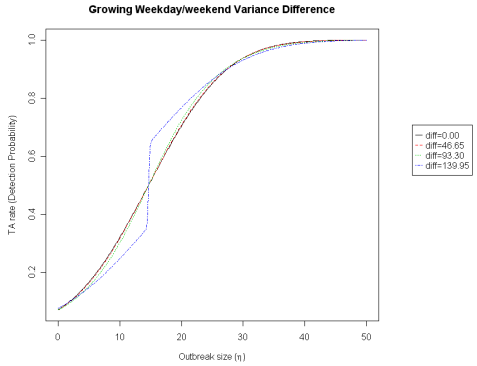

Fig. 5. Comparing Shewhart chart performance for forecasters with different residual seasonal variances (diff=difference between weekday and weekend residual variance) but identical overall variance $\sigma^2 = 100$.)

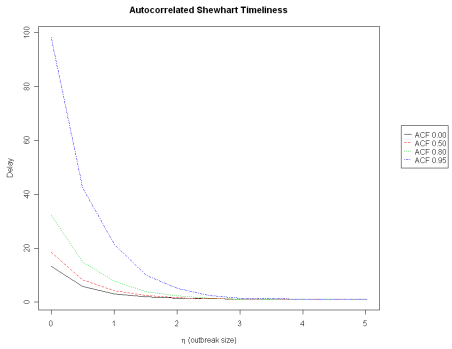

Fig. 6. Comparing Shewhart chart timeliness for forecasters with different residual autocorrelation levels (ACF) but identical overall variance $\sigma^2 = 1$.)

In conclusion, if there is strongly differentiated seasonal variance, an improved RMSE will not always give better detection performance, depending on the size of the outbreak. For some outbreak sizes, a forecaster with a larger overall RMSE but low weekend RMSE can outperform a forecaster 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 forecasters for each season.

V. Empirical Validation

We have shown theoretical results for detection and timeliness under different forecasters and outbreaks. We performed simulation experiments to examine the effects of autocorrelation, and to evaluate the applicability to real-world data.

A. 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 tests using spike outbreaks, no significant deviation was seen from the theoretical performance, when the control limit was set according to the final resulting variance. The detection performance is not affected by autocorrelation. However, figure 6 shows a significant deterioration in timeliness for small outbreak sizes and high autocorrelation. This agrees with Wheeler (9) regarding the relatively small impact of most autocorrelation levels on Shewhart charts.

B. Authentic Biosurveillance Data

An authentic health dataset is now used to determine the effectiveness of theory when estimating performance of currently-used forecasters. Three forecasters were compared: multiplicative Holt-Winters (triple exponential smoothing), regression (using trend, day-of-week indicators, and sin and cos terms for yearly seasonality), and 7-day differencing. Full descriptions of each of these methods can be found in (4). Simulated outbreak signals of various sizes were inserted on multiple possible test days, and the results for detection and timeliness were calculated. The RMSE for each forecaster was also computed and used to generate a theoretical performance curve for each forecaster.

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, and so a further examination was done, with seasonal variance computed for each day-of-week and performance predicted using seasonal variance computations. The results are in Figure 7, where an improved fit is seen, especially for the Holt-Winters residuals, though there is still some difference on the larger outbreaks.

Figure 8 compares the timeliness performance of real forecasters 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. The extra delay for regression and 7-day differencing seems to be due to autocorrelation (the residuals from these methods have significant residual autocorrelation, while Holt-Winters does not) but 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 forecasters' performance ranking is related to their RMSE ranking, as expected.

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

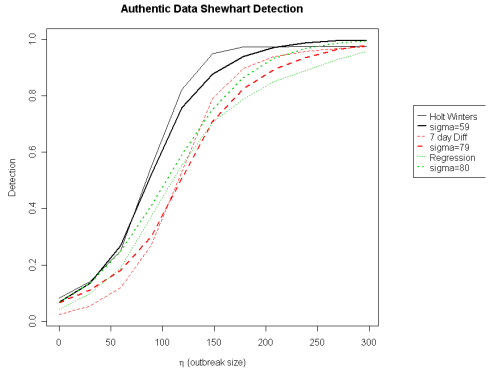

Fig. 7. Comparing actual (thin) and theoretical (thick) Shewhart chart performance for forecasters with different RMSEs, assuming day-of-week variance, as a function of outbreak size (η). Black/solid=Holt-Winters, Red/dashed=7-day Diff, Green/dotted=Regression.

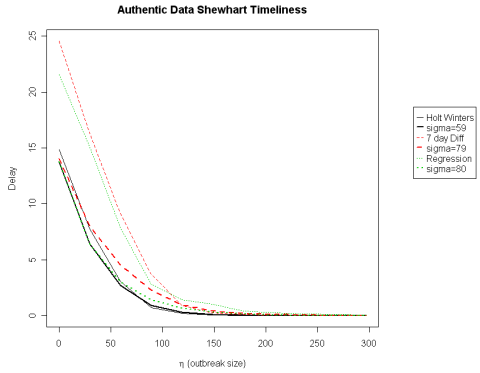

Fig. 8. Comparing actual (thin) and theoretical (thick) Shewhart chart timeliness for forecasters with different RMSEs, assuming constant variance, as a function of outbreak size (η). Black/solid=Holt-Winters, Red/dashed=7-day Diff, Green/dotted=Regression.

VI. 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 relationship between forecast precision and detection performance theoretically and quantified the effects under control chart assumptions. We have also exmained 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.

In conclusion, given the forecasting precision needed for useful detection, the question is whether that level of precision is possible. 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, to have a false alarm once every two weeks, and have a 95% chance of detecting a spike outbreak impacting 100 people, one would need normal residuals with a forecast RMSE < 32. In contrast, the best forecasting measures we have used here has an RMSE of 59. If we cannot accept a higher false alert rate, we must either find a way to further improve our forecasters, or tailor our detectors to specific outbreak signals.

Acknowledgment

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.

References

[1] S. E. Fienberg and G. Shmueli, "Statistical issues and challenges associated with rapid detection of bio-terrorist attacks," *Statistics in Medicine*, vol. 24(4), pp. 513–529, 2005.

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

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

[4] T. H. Lotze, S. P. Murphy, and G. Shmueli, "Preparing biosurveillance data for classic monitoring," *Advances in Disease Surveillance*, 2007.

[5] H. S. Burkom, S. P. Murphy, and G. Shmueli, "Automated time series forecasting for biosurveillance," *Statistics in Medicine*, vol. 26, pp. 4202–4218, 2007.

[6] H. D. Maragah and W. H. Woodall, "The effect of autocorrelation on the retrospective x-chart," *Journal of Statistical Computation and Simulation*, vol. 40, pp. 29–42, 1992.

[7] W. H. Woodall and F. W. Faltin, "Autocorrelated data and spc," *ASQC Statistics Division Newsletter*, vol. 13, pp. 18–21, 1993.

[8] C. S. Padgett, L. A. Thombs, and W. J. Padgett, "On the -risks for shewhart control charts," *Communications in StatisticsSimulation and Computation*, vol. 21, pp. 1125–1147, 1992.

[9] D. J. Wheeler, "Correlated data and control charts," in *Fifth Annual Forum of the British Deming Association*, 1992.