Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance

(to appear in *Quality Engineering*)

Ronald D. Fricker, Jr., Matthew C. Knitt, and Cecilia X. Hu
Naval Postgraduate School

February 25, 2008

Abstract

This paper compares the performance of two new directionally-sensitive multivariate methods, based on the multivariate CUSUM (MCUSUM) and the multivariate exponentially weighted moving average (MEWMA), for biosurveillance. While neither of these methods is currently in use in a biosurveillance system, they are among the most promising multivariate methods for this application. Our analysis is based on a detailed series of simulations using synthetic biosurveillance data that mimics various types of disease background incidence and outbreaks. We apply the MCUSUM and the MEWMA to residuals from an adaptive regression that accounts for the systematic effects normally present in biosurveillance data. We find that, much like the results from univariate CUSUM and EWMA comparisons in classical statistical process control applications, the directionally-sensitive MCUSUM and MEWMA perform very similarly.

1 Introduction

Biosurveillance is the process of monitoring health data in order to assess changes in disease incidence. Traditional biosurveillance methods have been focused on retrospectively analyzing medical and public health data, such as hospital admittance or mortality rates, to determine the existence of a disease outbreak (Shmueli, 2006) and/or to conduct epidemiological investigations (Stoto, 2007). Via traditional biosurveillance, the process of collecting and analyzing data can take days, even weeks, before definitively concluding that an outbreak has occurred and alerting officials. Improving on the timeliness of the alerts may help prevent spread of the disease and has the potential to significantly improve effective response to an outbreak or bioterrorism attack, particularly if the disease is contagious.

The Centers for Disease Control and Prevention (CDC) as well as many state and local health departments around the United States have started to develop and field electronic biosurveillance systems (CDC, 2004). Making use of existing health-related data, often already in electronic form, these surveillance systems are intended to give early warnings of bioterrorist attacks or other emerging health conditions. See Fricker et al. (2008), Fricker (2007a), Fricker (2007b), Stoto (2007), Fricker and Rolka (2006), Shmueli (2006) and Stoto et al. (2006) for more detailed discussions and related research.

Biosurveillance data frequently occurs in the form of discrete counts, such as daily counts of chief complaints at a hospital. Chief complaints are broad categories – e.g., respiratory, gastrointestinal, unspecified infection, or neurological – into which patients are grouped before diagnosis. Chief complaint is the primary symptom or reason a patient sought care. Biosurveillance might also be conducted on counts of positive lab test results, sales of over-the-counter medical products, and calls to emergency call centers, for example.

Biosurveillance data is also frequently autocorrelated with seasonal cycles (e.g., the flu season), trends (e.g., population changes), and other uncontrollable systematic features of the data. Unlike in the traditional SPC setting, in which it is generally reasonable to assume that under in-control conditions the data is independent and the distribution of the statistic being monitored is stationary, this is generally not the case with biosurveillance data. Thus, if standard SPC methods were applied to the raw data, signals would sometimes have an excessively high probability of occurrence even when the process was "in-control" and, similarly, there would be other times when the it would be excessively hard to signal even when the data was in an "out-of-control" state. This often leads to the need to first model the data, in order to appropriately adjust for or remove the known systematic features, and then to evaluate the "preconditioned data" or the model residuals for evidence of an outbreak. (See Lotze, Murphy, and Shmueli, 2006, for further discussion.)

Current biosurveillance systems run multiple simultaneous univariate statistical process control (SPC) procedures, each focused on detecting an increase in a single dimension. Multiple simultaneous univariate procedures have the advantages of ease of implementation and interpretation, though they have the potential to be less sensitive to some types of changes when compared to multivariate methods. This paper compares two new directionally-sensitive multivariate methods based on the multivariate CUSUM (MCUSUM) and the multivariate exponentially weighted moving average (MEWMA). While neither of these methods is currently in use in a biosurveillance system, they are among the most promising temporal multivariate methods for this application.

Rogerson and Yamada (2004) evaluated multiple univariate CUSUMs versus a directionally invariant multivariate CUSUM for monitoring changes in spatial patterns of disease. Recent work on directional multivariate procedures includes Joner et al. (2008); Fricker (2007a); Stoto et al. (2006); and, building on the work of Follmann (1996), Perlman (1969), and Kudô (1963), Testik and Runger (2006). For a review of the use of SPC applications and methods in the context of public health surveillance, see Woodall (2006).

The paper is organized as follows. In Section 2 the MCUSUM and MEWMA procedures are described, including how they were applied to residuals from an adaptive regression-based model. Section 3 describes how we generated synthetic background disease incident counts and outbreaks, and Section 4 describes the comparison methodology, including how we determined the form of the adaptive regressions used and how we selected various parameter values for the MCUSUM and MEWMA procedures. Section 5 presents the results of the simulation comparisons and then illustrates the application of the methods on actual biosurveillance data from five hospitals located in a large metropolitan area. The paper concludes in Section 6 with a discussion of the implications of our findings and some recommendations.

2 Methods

In this section we first describe the MCUSUM and the MEWMA with a focus on the directionally-sensitive variants that are relevant to the biosurveillance problem. We then describe the "adaptive regression with sliding baseline" approach of Burkom et al. (2006). The use of adaptive regression is motivated by the need to remove systematic trends commonly present in biosurveillance data and the MCUSUM and MEWMA are subsequently run on one-day ahead (standardized) forecast errors which we assume are continuous.

2.1 Directional MCUSUM

Consider a p-dimensional set of observations at time t, $\mathbf{X}_t = \{X_1, \ldots, X_p\}$. Crosier (1988) proposed a MCUSUM that at each time t calculates the statistic

$$\mathbf{S}_t = (\mathbf{S}_{t-1} + \mathbf{X}_t - \boldsymbol{\mu})(1 - k/d_t), \text{ if } d_t > k, \tag{1}$$

where $\boldsymbol{\mu}$ is the mean of $\mathbf{X}_t$, k is a predetermined statistical distance, and $d_t = [(\mathbf{S}_{t-1} + \mathbf{X}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{S}_{t-1} + \mathbf{X}_t - \boldsymbol{\mu})]^{1/2}$. If $d_t \leq k$ then reset $\mathbf{S}_t = \mathbf{0}$. The procedure starts with $\mathbf{S}_0 = \mathbf{0}$ and sequentially calculates

$$C_t = (\mathbf{S}_t'\boldsymbol{\Sigma}^{-1}\mathbf{S}_t)^{1/2},$$

where $\mathbf{\Sigma}$ is the variance-covariance matrix of $\mathbf{X}$. It concludes that a change has occurred at the first time when $C_t > h$, for some pre-specified threshold h that achieves a desired average time between false signals (ATFS). See Section 4 for a discussion of the ATFS metric.

In terms of choosing $\mathbf{k}$, Crosier (1988) states, "In the univariate [CUSUM] case, the quantity $S_{t-1} + (X_t - \mu)$ is shrunk towards 0 by k standard deviations. If this is to hold for the multivariate case, $\mathbf{k}$ must satisfy $\mathbf{k}'\mathbf{\Sigma}^{-1}\mathbf{k} = k^2$ – that is, $\mathbf{k}$ must be of length k, where the length is defined by using the covariance matrix $\mathbf{\Sigma}$."

The literature contains a number of MCUSUM procedures. In fact, the Crosier procedure described above is one of a number of other multivariate CUSUM-like algorithms he proposed, but Crosier generally preferred the above procedure after extensive simulation comparisons. Pignatiello and Runger (1990) proposed other multivariate CUSUM-like algorithms but found that they performed similar to Crosier's. Healy (1987) derived a sequential likelihood ratio test to detect a shift in a mean vector of a multivariate normal distribution. However, while Healy's procedure is more effective when the change is to the precise mean vector to be detected, it is less effective than Crosier's for detecting other types of shifts, including mean shifts that were close to but not precisely the specified mean vector.

For the biosurveillance problem, an advantage of Crosier's MCUSUM formulation is that it is easy to modify to only look for positive increases. As described in Fricker (2007a), the motivation for this modification is the univariate CUSUM where directionality is achieved because the CUSUM statistic is bounded below by zero. In the modified MCUSUM directionality is similarly achieved by bounding each component of the cumulative sum vector by zero. In particular, for detecting positive increases relevant to the biosurveillance problem, when $d_t > k$ limit $\mathbf{S}_t$ to be non-negative in each dimension by replacing Equation (1) with $\mathbf{S}_t = (S_{t,1}, \ldots, S_{t,p})$ where

$$S_{t,j} = \max[0, (S_{t-1,j} + X_{t,j} - \mu_j)(1 - k/d_t)],$$

for $j = 1, 2, \ldots, p$.

2.2 Directional MEWMA

Lowry et al. (1992) introduced the MEWMA as a generalization of the univariate EWMA of Roberts (1959). As with the MCUSUM, denote the mean for $\mathbf{X}_t$ as $\boldsymbol{\mu}$ and let $\mathbf{\Sigma}$ be the covariance matrix. In the spirit of the reflected EWMA of Crowder and Hamilton (1992), the directionally sensitive MEWMA proposed by Joner et al. (2008) calculates

$$\mathbf{Z}_t = \begin{cases} \max[\mathbf{0}, \lambda(\mathbf{X}_t - \boldsymbol{\mu}) + (1-\lambda)\mathbf{Z}_{t-1}], & \text{for } t > 0 \\ \mathbf{0}, & \text{for } t = 0 \end{cases},$$

where the maximum function is applied componentwise. $\mathbf{Z}_t$ is a weighted average of the current observation standardized around $\mathbf{0}$ and the previous $\mathbf{Z}$ statistic. The parameter $0 < \lambda \leq 1$ is the *smoothing parameter* which controls the weight assigned to the new observation vector. The covariance matrix for $\mathbf{Z}_t$ is

$$\boldsymbol{\Sigma}_{\mathbf{Z}_t} = \frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{2-\lambda}\boldsymbol{\Sigma}.$$

Taking the limit as $t \to \infty$, we have

$$\boldsymbol{\Sigma}_{\mathbf{Z}_\infty} = \frac{\lambda}{2-\lambda}\boldsymbol{\Sigma}.$$

$\boldsymbol{\Sigma}_{\mathbf{Z}_\infty}$ is then used to calculate the MEWMA test statistic E_t where

$$E_t = \mathbf{Z}_t'\Sigma_{\mathbf{Z}_\infty}^{-1}\mathbf{Z}_t.$$

The MEWMA signals an alarm whenever E_t exceeds a predetermined threshold h which is set to achieve a desired ATFS. If E_t does not exceed h, then the MEWMA iterates through the next time step with a new observation vector, recalculating the test statistic, and continuing until such time as the $E_t > h$.

2.3 Adaptive Regression with Sliding Baseline

We used the “adaptive regression model with sliding baseline” of Burkom et al. (2006) to model and, to the greatest extent possible, remove the systematic components of biosurveillance data. The basic idea is as follows. Let Y_i be an observation, say chief complaint count on day i at one of p hospitals. For each hospital, regress the observations from the past n days on time relative to the current period. Then use the model to predict today’s observation and apply the MCUSUM or MEWMA to the vector of the differences between today’s observed value and the predicted value. Repeat this process each day, always using the most recent n observations as the sliding baseline in the regression to calculate the forecast error.

For $t > n$, and assuming a simple linear formulation, the model for each hospital is

$$Y_i = \beta_0 + \beta_1 \times (i - t + n + 1) + \epsilon \qquad (2)$$

for $i = t-1, \ldots, t-n$. Of course, as appropriate, the model can also be adapted to allow for nonlinearities by adding a quadratic term or to allow for day-of-the-week effects by including the appropriate indicator variables in Equation (2).

Burkom et al. (2006) used an 8-week sliding baseline ($n = 56$). We compared the performance for a variety of ns and between a linear and quadratic form of the model. Section 4.1 describes how we determined the form for the adaptive regression and the length of the sliding baseline.

The model is fit using ordinary least squares, regressing $Y_{t-1}, \ldots, Y_{t-n}$ on $n, \ldots, 1$. Having fit the model, the forecast error is

$$r_t = Y_t - \left[\hat{\beta}_0 + \hat{\beta}_1 \times (n+1)\right],$$

where $\hat{\beta}_0$ is the estimated intercept and $\hat{\beta}_1$ is the estimated slope. If we denote the forecast error for hospital j as $r_t(j)$, then the residual vector upon which the MCUSUM and MEWMA are run is $\mathbf{X}_t = \{r_t(1)/\sigma_Y(1), \ldots, r_t(p)/\sigma_Y(p)\}$, where $\sigma_Y(j)$ is the standard deviation of the dependent variable in the adaptive regression.

3 Simulating Biosurveillance Data

In order to compare the methods, we simulated a background disease incidence and then overlaid various types of simulated bioterrorism attacks/natural disease outbreaks (which we will refer to herein simply as "outbreaks"). The simulations were conducted in MatLab 7.1.0.246 using the *randn* function to generate normal random variates. The simulations of both background disease incidence and outbreaks are purposely idealized depictions designed to capture the main features of biosurveillance data. The use of simulation and the idealization of the data features were done for two very specific reasons:

- So that we could definitively compare and contrast the relative performance of the various procedures under known conditions, and
- So that we could clearly distinguish how the various features of the data did or did not affect each procedures' performance.

The background disease incidence data was simulated as the sum of a mean disease incidence, a seasonal sinusoidal cycle, and a random fluctuation. Outbreaks, when they occurred, were incorporated as another additive term. That is, a daily observation Y_t was simulated as

$$Y_t = \max(0, \lceil m + s_t + o_t + N \rceil), t = 1, 2, 3, \ldots, \qquad (3)$$

where

Scenario	m	A	σ
1	90	80	30
2	90	80	10
3	90	20	30
4	90	20	10
5	90	0	30
6	90	0	10

Table 1: Parameters for Equation (3) for scenarios 1-6.

- m is the annual mean level of disease incidence;
- s_t is the seasonal deviation from the mean, calculated as $s_t = A[\sin(2\pi t/365)]$, where A is the maximum deviation from m with $t = 1$ corresponding to October 1st on a 365 day per year calendar;
- o_t is the mean outbreak level which, when an outbreak is occurring, increases the disease incidence level as described below;
- N is the random noise around the systematic component, modeled as $N \sim N(0, \sigma)$ independently in each dimension; and,
- $\lceil x \rceil$ is the ceiling function, which rounds x up to the next largest integer.

Table 1 specifies the parameter values for Equation (3) which define six "scenarios" designed to span a range of possible underlying disease incidence patterns with large counts. The parameters were selected to generate synthetic data that mimics disease incidence patterns similar to selected data sets at the CDC's EARS simulation data sets (www.bt.cdc.gov/surveillance/ears/datasets.asp). In particular, $m = 90$, $A = 80$, and $\sigma = 30$ or $\sigma = 10$ result in disease incidence patterns similar to EARS data set S08. Setting $m = 90$, $A = 20$, and $\sigma = 10$ results in disease incidence patterns similar to the S01 data set, as well as other patterns that are intermediate between S01 and S08. The specific EARS data sets that we mimicked were chosen in consultation with a CDC expert (Hutwagner, 2006).

Within each scenario we simulated four streams of data, representing say the chief complaint counts from four large hospitals or perhaps the aggregate counts from four metropolitan areas for one type of syndrome. Various combinations of A and σ result in various covariances between the data streams. For example, $A = 80$ and $\sigma = 10$ result in $\rho = 0.97$ while $A = 80$ and

$\sigma = 30$ result in $\rho = 0.78$. In comparison, $A = 20$ and $\sigma = 10$ result in $\rho = 0.68$ while $A = 20$ and $\sigma = 30$ result in $\rho = 0.19$. And, of course, $A = 0$ with either $\sigma = 30$ or $\sigma = 10$ results in $\rho = 0$.

While these choices may seem either arbitrary or too restrictive, meaning they do not characterize some particular pattern that occurs in a particular biosurveillance setting, we chose them because they capture a wide range of data. Furthermore, as will be shown in Section 5, and as was also demonstrated in Fricker et al. (2008) and Dunfee and Hegler (2007), the adaptive regression turns out to be remarkably good at removing the systematic trends in the data so that the specific choices made above are actually of little import and have little impact on the final result.

Outbreaks were incorporated into Equation (3) as an additive term $o(t)$ representing the mean outbreak level, parameterized in terms of a peak magnitude M, a duration D, and a random start day τ. Outbreaks increased linearly up to M and then linearly back down to zero:

$$o(t) = \begin{cases} M\left[2(t-\tau+1)/(D+1)\right], & \tau \le t \le \tau + D/2 - 1/2 \\ M\left[1-(2(t-\tau)-D+1)/(D+1)\right], & \tau + D/2 - 1/2 < t \le \tau + D - 1 \\ 0, & \text{otherwise.} \end{cases}$$

We evaluated the procedures' performance for outbreaks of various magnitudes and durations. We used three magnitudes – small, medium, and large – defined as a fraction of the annual mean disease incidence m: $M = 0.1m = 9$, $M = 0.25m = 22.5$, and $M = 0.5m = 45$, respectively. For all the scenarios we looked at durations that ranged from short to long: $D = 3, 5, \ldots, 15$ days. In the simulations, τ was the same in all dimensions, as was M and D. So, when an outbreak occurred, it occurred at the same time and equally across all hospitals.

As we previously mentioned, the characterization of disease incidence in Equation (3) is purposely idealized in order to facilitate comparison of the relative performance of the procedures under various scenarios. The idea is to mimic the most salient and important features of biosurveillance data in a simulation environment where we can know precisely when outbreaks occur so that we can clearly assess and evaluate performance. That said, it is important to note that the methods do not exploit the idealized features of the data and can be readily adapted to account for those features of real data that are not included in Equation (3). For example:

1. *Regular seasonal cycles.* The seasonal cycle in Equation (3) is idealized and could be exploited to make artificially accurate predictions. That is, not only do the cycles occur at precisely the same time each year, but they are perfect sinusoids and they are synchronized

across all data streams. However, since each data stream is modeled separately and because the length of time over which the adaptive regressions are run is too short to model the cycle (see Section 4.1), the MCUSUM and MEWMA procedures do not use the information in the idealized seasonal cycle.

2. *No linear trends.* Growing or shrinking populations, or changes in health conditions, could result in linear (or other) trends in the disease incidence. A trend term is not included in Equation (3) since, if the procedures can appropriately adjust for the seasonal component, they can also adjust for linear trends.

3. *No day-of-the-week, holidays or other such effects.* Dunfee and Hegler (2007) demonstrated that adding day-of-the-week effects into Equation (3) made little difference in the performance of the adaptive regression given a sufficiently long sliding baseline (roughly $n > 30$ days or so). Hence we neither simulate day-of-the-week and other such effects as an unnecessary complication that does not affect the results or conclusions. (However, we do illustrate the application of the methods on actual biosurveillance data in Section 5, the data for which do contain day-of-the-week effects.)

See Shmueli (2006), Lotze et al. (2006), and Burkom et al. (2006) for detailed expositions on the features of biosurveillance data. See Fricker et al. (2008) and Dunfee and Hegler (2007) for examples of how the adaptive regression methodology was able to account for and remove day-of-the-week effects from synthetic biosurveillance data. See Kleinman et al. (2005) for an alternate methodology designed to simulate biosurveillance data in both space and time.

4 Comparison Methodology

The metrics used to compare performance between procedures were: (1) the fraction of times a procedure missed detecting an outbreak and (2) the average time to first outbreak signal (ATFOS). The former is a measure of detection capability while the latter is a conditional measure of the timeliness of detection. The ATFOS is defined as the average time until the first signal among all simulations for which a signal occurred during the outbreak period. Clearly performance in both dimensions must be considered since a desirable procedure must simultaneously have a short ATFOS and a low fraction of outbreaks missed. A procedure that is small in one dimension while being large in the other is not particularly useful.

This approach differs from much of the biosurveillance literature that attempts to evaluate performance simultaneously in three dimensions: "sensitivity, specificity, and timeliness." While we do assess performance in terms of timeliness via ATFOS and fraction missed, a sensitivity-like measure, we use a fixed average time between false signals (ATFS) in the third dimension to simplify the analysis. In so doing, we assume that the relative performance of the procedures does not change for other choices of ATFS.

This is similar to the approach used in the statistical process control (SPC) literature, where the ATFS is roughly equivalent to the "in-control average run length" and the ATFOS is equivalent to the "out-of-control average run length." The average run length, or ARL, is the average number of observations until a signal. In the SPC literature, it is the common and well accepted practice to compare the performance of procedures by first setting thresholds that achieve a specific in-control average run length and then compare out-of-control average run lengths under various conditions. The procedure that demonstrates lower out-of-control average run lengths across a variety of conditions deemed important is judged to be the better procedure.

However, this approach differs from the SPC literature because we also use the fraction missed metric. In the SPC literature, once a process goes out-of-control, it is assumed to stay in that condition until a procedure signals and the cause is identified and corrected. Thus, once a procedure goes out of control, any signal is a true signal. This is not the case in biosurveillance where outbreaks are transient and after some period of time disappear. In this situation, it is possible for a procedure to fail to signal during an outbreak, after which a signal is a false signal.

Returning to the biosurveillance literature's "specificity" metric, we prefer ATFS because the concept of specificity is not well defined in sequential testing problems. In classical hypothesis testing, specificity is the probability that the null hypothesis is not rejected when it is true. It is one minus the probability of a Type I error. In medicine, it is the probability that a medical test will correctly indicate that an individual does not have a particular condition. However, biosurveillance involves sequential testing where, in the absence of an outbreak the repeated application of any procedure will eventually produce a false signal. Said another way, in the absence of an outbreak, one minus the specificity for a sequential test must approach 100 percent as the number of tests is allowed to get arbitrarily large.

In the biosurveillance literature, specificity is often (re)defined as the fraction of times a procedure fails to signal divided by the number of times the procedure is applied to a stream of biosurveillance data without outbreaks (c.f., Reis et al. 2003). If the data is independent and identically distributed from day to day, and if the test procedure results in test statistics that are independent and identically distributed from day to day as well, then such a calculation is an appropriate estimate of the specificity of a test on a given day. However, biosurveillance data are generally autocorrelated and, even if it were not, any procedure that uses historical data in the calculation of a test statistic will produce autocorrelated statistics. Under these conditions, it is not clear what the quantity from the above calculation represents. It is certainly not specificity in the classical hypothesis testing framework. See Fraker et al. (2007) for additional discussion.

So, for each scenario in Table 1, we determined the threshold for each procedure that gave an ATFS of 100 days. The ATFS is a measure of the time between clusters of false signals. It would be equivalent to the average time between signal events (ATBSE) metric of Fraker et al. (2007) if they allowed the procedure to be reset to its initial condition after a "signal event" and the data related to the signal event is removed from the adaptive regression. All other things being equal, larger ATFS values are to be preferred.

The thresholds to achieve a particular ATFS were determined empirically as follows. For a particular scenario and procedure, we chose an initial h and ran an algorithm r times (starting with $\mathbf{S}_0 = \mathbf{0}$ or $\mathbf{Z}_0 = \mathbf{0}$), recording for each run i the time t_i when the procedure first signalled. The ATFS was estimated as $\sum_{i=1}^{r} t_i/r$, and we then iteratively adjusted h and re-ran the procedure to achieve the desired ATFS, eventually setting r large enough to make the standard error of the estimated ATFS acceptably small (less than one day). Once the thresholds were set, the procedures were then compared across all the scenarios specified in Table 1 for all the outbreak types described in Section 3.

The purpose of setting the thresholds to achieve equal time between false alarms was to ensure a fair comparison between the procedures. That is, it is always possible to improve a procedure's ability to detect an actual outbreak by lowering the threshold, but this comes at the expense of also decreasing the ATFS. Thus, by first setting the thresholds to achieve equal time between false alarms we could then make an objective judgement about which procedure or procedure was best at detecting a particular type of outbreak.

Across all the scenarios in Table 1, the MCUSUM thresholds ranged from $h = 3.25$ to $h = 3.31$. For the MEWMA procedures, the thresholds ranged from $h = 4.57$ to $h = 4.78$. The variation is due to differences in the lengths of the sliding baseline for the adaptive regressions used for each scenario and the resulting ability of the adaptive regressions to remove the systematic effects from the data. In the absence of a systematic component in the data (i.e., scenarios 5 and 6), we set $h = 3.25$ for the MCUSUM and $h = 4.6$ for the MEWMA.

Having set the thresholds to achieve equal ATFS performance, the ATFOS and fraction missed were calculated as follows. For each iteration i, the procedures were run for 100 time periods (using data from $100 + n$ time periods so that the adaptive regression could be fit for period 1) without any outbreaks. If a procedure signalled during this time it was reset and restarted, just as it would be in a real application, since the signal corresponded to a false alarm. This allowed the Z_{100} and S_{100} statistics to be in a steady state condition at the time of the outbreak. Outbreaks began at time 101 and continued for the appropriate duration. If the procedure signalled at time t_i within the duration of the outbreak, the time to first outbreak signal was recorded as $t_i - 100$ and the ATFS was estimated as $\sum_{i=1}^{s}(t_i - 100)/s$ for the s iterations that signalled within the outbreak duration. The fraction missed was calculated as the number of iterations for which the procedure failed to signal during the outbreak divided by the total number of iterations run.

4.1 Determining the Form of the Adaptive Regressions

When using regression to predict future observations, the question naturally arises as to how much historical data should be used for the regression's sliding baseline. Of course, all other factors being equal, regressions based on a shorter sliding baseline will less accurately estimate the underlying systematic trends in the data than those based on longer sliding baselines. However, while a longer sliding baseline should allow for a more detailed regression model and presumably a better prediction, often in biosurveillance the amount of available data is limited or the older data of questionable relevance due to changing trends or phenomena. Hence, there is a trade-off to be made between the amount of historical data used in a particular model and the predictive accuracy of that model.

As described in Dunfee and Hegler (2007), this led us to also determine and evaluate the performance of the "optimal" sliding baseline (n) for each scenario (m, A, σ combination). For each of the six scenarios we studied, we determined the optimal n to later use in the actual regression analysis and method comparisons. In addition, two separate regression models were evaluated in order to determine the best form of the model, either linear or quadratic.

Figure 1 shows an example of how we assessed the form of the adaptive regression and determined the "optimal" sliding baseline for scenario 1 ($m = 90, A = 80, \sigma = 30$). The optimal

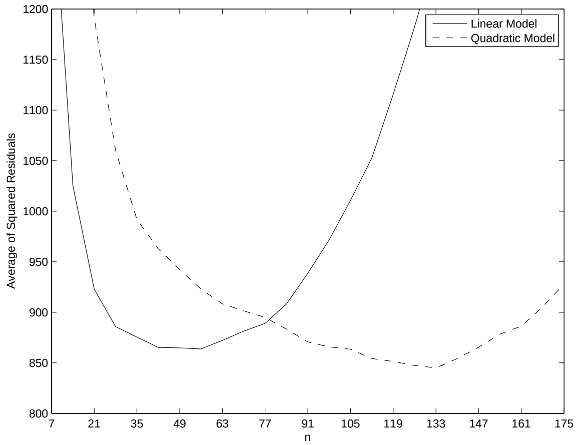

Figure 1: Average squared residuals for linear and quadratic models as a function of the amount of historical data – i.e., the size of the sliding baseline (n) – used to fit the regression models under scenario 1. From this, we determined the best model was linear with an "optimal n" of about 30 days.

n was chosen by visual inspection with the criteria that the n be as small as possible but also as close to achieving the minimum average squared residual as possible. This means that we chose the smallest n that achieved most of the reduction in the average squared residual and not the n that occurred precisely at the minimum point on the curve. For example, in Figure 1 we determined that the "optimal n" was about 30 days for the linear model (and much larger for the quadratic model).

Figure 1 shows that the linear model achieved close to the same minimum average squared residual as the quadratic model at a much smaller n. As described in Dunfee and Hegler (2007), this occurred consistently for all of the scenarios leading us to choose a linear adaptive regression model in all of our evaluations. For the linear model, across all the scenarios, the optimal ns ranged from 30 to 45 days. For other scenarios with day-of-the-week effects, not described here (see Fricker et al., 2008), the optimal ns were even larger with the largest being around 56 days – the size recommended by Burkom et al. (2006).

4.2 Determining λ for the MEWMA and k for the MCUSUM

In order to compare the MEWMA and MCUSUM under a variety of biosurveillance scenarios, we wanted to first set their parameters such that they performed as similarly as possible under the standard statistical process control assumptions of *iid* observations and a sustained jump change in the mean. To do this, we first fixed λ for the MEWMA and then searched for the value of k in the MCUSUM that matched its performance to the MEWMA's.

Setting λ is a trade-off between how much emphasis is put on past information in the MEWMA statistic and the desire for the MEWMA to be as sensitive as possible to changes in $\mathbf{X}$. The idea is that, in both the univariate and multivariate EWMA, larger λs put more weight on the current observation and less on past observations. At its most extreme, setting $\lambda = 1$ turns the univariate EWMA into a Shewhart procedure. Montgomery (2001) recommends setting $0.05 \leq \lambda \leq 0.25$ for the univariate EWMA and, given the emphasis on timeliness in this application and based on our experience (c.f. Chang and Fricker, 1999), we thus chose to set $\lambda = 0.2$.

Having fixed λ, we conducted simulation comparisons over various values of k to find that value for which the MCUSUM performed as closely as possible to the MEWMA. We did this by comparing how well the MCUSUM detected various sustained mean shifts for a four dimensional standard multivariate normal. As shown in Figure 2, we found that $k = 0.74$ gave the closest performance to the MEWMA with $\lambda = 0.2$, where we set all of the components of the $\mathbf{k}$ vector equally (e.g., for $k = 0.74$ we set $\mathbf{k} = \{0.37, 0.37, 0.37, 0.37\}$).

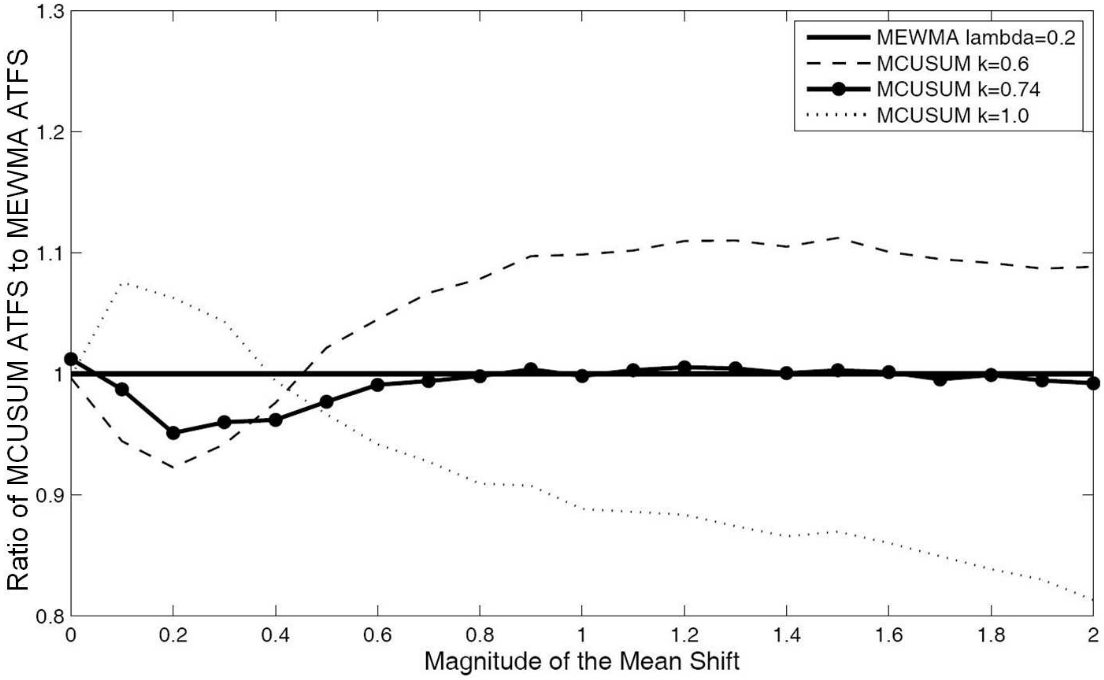

Figure 2: Comparison of the MCUSUM for various k to the MEWMA with $\lambda = 0.2$ in terms of the ratio of the MCUSUM ATFS to the MEWMA ATFS. The MCUSUM with $k = 0.74$ came closest to matching the MEWMA's performance over a wide range of mean shifts.

In particular, Figure 2 shows the percent change in the ATFS (which, in this case, is the same as the average run length measure, or ARL, used in the traditional statistical process control literature) for the MCUSUM for various k compared to the MEWMA. In the plot we see that the MCUSUMs with $k < 0.74$ signal faster than the MEWMA for small mean shifts and signal slower as the mean shift increases. Conversely, as k increases from 0.74, the MCUSUM begins to signal slower than the MEWMA for small shifts and much faster as the mean shift increases. Ultimately, we found that the MCUSUM with $k = 0.74$ had the closest performance over a wide range of shifts: it achieves almost precisely the same ATFS for shifts between about 0.6 and 2.0, and is at most about five percent off over the entire range of shifts considered, from 0 to 2.0.

5 Results

In this section we compare the MCUSUM to the MEWMA, first using the simulated biosurveillance data and then on actual data. In both cases we use simulated or real data that consist of daily counts.

5.1 Simulation Comparison Results

Figures 3 and 4 summarize our main finding: the MEWMA and MCUSUM performed virtually identically, both in terms of ATFOS and percent missed, across all the scenario and outbreak combinations we evaluated. Though the lines deviate slightly in Figures 3 and 4, the differences are not statistically significant. See Hu and Knitt (2007) for details.

Specifically, in Figure 3 we see that there is no difference in MCUSUM and MEWMA performance for scenario 4 across all the types of outbreaks, from small to large magnitudes and for all the durations. This result was also true for the other five scenarios. For example, Figure 4 shows the results for scenarios 1, 3, and 5 for an outbreak of medium magnitude. See Hu and Knitt (2007) for plots for all of the scenarios and types of outbreaks.

Figure 3 demonstrates how the procedures perform for the various types of outbreaks. For example, the ATFOS plots show that outbreaks of small magnitude and of three days duration will only be detected about 30 percent of the time and, when detected, it will take about two days on average for either the MCUSUM or MEWMA to signal. As the outbreak magnitude increases, the procedures detect virtually all of the outbreaks and the ATFOS decreases to about one day for the largest magnitude outbreak. In comparison, for durations of 15 days, the methods detect almost 70 percent of the small magnitude outbreaks and again virtually all of the larger outbreaks. For the small magnitude outbreaks the average time to signal is about six days, for the medium magnitude it is just under five days, and for the large magnitude outbreak it is about 2-1/2 days.

Figure 4 demonstrates that the adaptive regression with sliding baseline methodology does very well at removing the systematic component, at least for our synthetic biosurveillance data. In this case, the systematic component is the seasonal sinusoid where, at the top the sinusoid is large ($A = 90$), in the middle it is medium sized ($A = 20$), and at the bottom it is non-existent ($A = 0$). In terms of ATFOS, there is no visible difference between the three plots in Figure 4. In terms of percent of outbreaks missed, there is a slight degradation in the number of outbreaks caught as the amplitude increases. However, these plots demonstrate that, overall, the adaptive regression is quite effective at accounting for the systematic trends in the data.

5.2 A Comparison Using Actual Biosurveillance Data

In this section we compare how the procedures perform on actual biosurveillance data using respiratory chief complaint counts for five hospitals located in one metropolitan area from October 1, 2001 to March 31, 2004. We focus on respiratory chief complaint data since it tends to include those patients who come to emergency rooms with the flu and flu-like symptoms, which could also be leading indicators of certain bioterrorism agents (Espino, 2001).

As originally described in Fricker (2007a), Figure 5 shows the respiratory chief complaint count data by hospital with a four week centered moving average overlaid. A number of features of the data are clear from the figure, including:

- The hospital moving averages do not exhibit increasing or decreasing trends, indicating the long-term incidence rate for respiratory chief complaints is relatively constant.
- There are significant differences in mean counts between hospitals, indicating that some hospitals either serve larger populations or serve populations with higher respiratory illness rates (or both), as well as significant variation in the raw counts around the smoothed mean.
- The hospital counts are positively correlated where, using the first six months of the data, the correlations between all pairs of hospitals lie in the interval $0.0 \leq r \leq 0.49$.

- There are visible "events" in the data that persist for periods of time. For example, there are peaks across most or all of the hospitals in March-June 2003 and December 2003-January 2004 that likely correspond to flu outbreaks.

The visible "events" are most likely naturally occurring flu outbreaks. They are consistent with the CDC's aggregate data on "percentage of visits for influenza-line illness reported by sentinel physicians." Specifically, the CDC (2005) characterized flu incidence in the South Atlantic region of the United States (where the hospitals are located) as follows.

- The 2002-2003 flu season was "mild" with the percentage of visits in the South Atlantic region peaking in February-March 2003. "Sporadic activity" was also reported in April and May 2003.

- The 2003-2004 flu season "began earlier than most seasons and was moderately severe." The percentage of visits in the South Atlantic region peaked in December 2003.

5.2.1 Implementation

To implement the procedures, we divided the data up into a "historical" set of data, consisting of the first six months (10/1/01-3/31/02), and a "future" set of data – the remaining two plus years (4/1/02-5/17/04). As one would do in practice, the idea was to use the "historical" data to estimate various quantities necessary for the procedures and then to illustrate each procedures' performance on the "future" data.

Following the approach of Burkom (2006), we transformed the data using the started log and then used the first eight weeks of data (10/1/01-11/25/01) as the initial sliding baseline for an adaptive regression. Given that day-of-the-week effects were present in the data, the adaptive regression was appropriately parameterized to include indicator variables in order to account for such effects. The regression was used to predict the started log transformed counts for day 57 and to calculate the resulting residuals. We then incremented the sliding baseline, calculated the prediction residuals for day 58, and continued on through the remainder of the first six months of the data, calculating the prediction residuals through day 182 (3/31/02).

From this, we used the prediction residuals from 11/26/01 to 3/31/02 to estimate the standard deviations of the started log transformed counts for each of the hospitals, which we then used to standardize the residuals and to estimate the variance-covariance matrix of the standardized residuals. For the "future" data, the standard deviations were also used to standardize the future adaptive regression residuals and the estimated variance-covariance matrix was used in the two procedures as described in Section 2.

The remaining parameters for the MCUSUM and MEWMA were determined based on the results of the previous simulations: we set $k = 0.74$ for the MCUSUM and $\lambda = 0.2$ for the MEWMA. For the thresholds, as in the simulations, we set $h = 4.6$ for the MCUSUM and $h = 3.25$ for the MEWMA and in order to achieve an ATFS of approximately 100 days.

5.2.2 Performance Comparisons

Figure 6 displays the signal times for the MEWMA and MCUSUM when they are run on the "future" respiratory data. The figure shows the smoothed means (of Figure 5) and first signal times overlaid. ("First signal time" means that repeated signals within 60 days after the first signal are suppressed for plot clarity.) What this figure shows is that the modified MCUSUM and the MEWMA performed extremely similarly. Specifically, seven of the eight first signal times occurred on exactly the same day for both procedures. The only signal that differed was the one on September 24th, where the MEWMA signaled on that day and the MCUSUM signaled on September 25th.

Furthermore, not shown on the plot are the other signals, of which 24 matched again to the day, three differed by only one day, and in only once case did one procedure (the MEWMA) signal without the other signaling. The net result is that the two procedures performed almost identically, with the small observed differences potentially attributable to the imprecision in the setting of the procedures' parameters and thresholds. Simply put, there was no practical difference, and almost no quantitative difference, in the performance of the MCUSUM and the MEWMA.

Figure 6 also gives some insight into how the procedures would perform in the real world. Specifically, we see that the methods do detect the flu outbreaks (the gray shading highlight the flu periods that the CDC identified for the South Atlantic region of the U.S.) with signals on February 17, 2003 and December 20, 2004. The former is consistent with an up-tick in the counts for the top hospital and the latter for up-ticks in four of the five hospitals. Also, the signal on September 24, 2003 is consistent with up-ticks in the counts of at least two of the five hospitals, even though this was not identified as a flu period. The other five signals are not related to obvious up-ticks in the counts, though that is certainly subject to interpretation.

However, even if they are false signals, they are within the expected number of false signals based on an ATFS of 100 days.

6 Discussion

In this section we summarize our findings on the MEWMA and MCUSUM comparisons and discuss some thoughts about applying these and similar methods to the practice of biosurveillance that arose during the course of the research.

6.1 MCUSUM or MEWMA?

When we began this research we fully expected to identify scenarios in which the MCUSUM performed better than the MEWMA and vice versa. That the two procedures performed practically identically is an unexpected surprise. It is a surprise because, while it is well-known that with the appropriate choice of parameters the univariate EWMA and CUSUM can be made to perform similarly in standard SPC applications, the directional MEWMA and MCUSUM described herein are neither the exact multivariate analogues of their univariate counterparts nor is the biosurveillance problem the same as the standard SPC application.

Because there is seemingly no performance advantage in using one method over the other, this result leads us to prefer the MEWMA for procedural reasons. Specifically, it is relatively easy to develop an intuitive appreciation for how to choose λ and much more difficult to understand how to appropriately choose $\mathbf{k}$. That is, unlike the k in the univariate CUSUM which has a clear interpretation, namely it is one-half of the smallest mean shift that is to be detected quickly, the k in Crosier's MCUSUM is a parameter in a multiplicative "shrinkage factor" for which there is no literature or research to guide one in the trade-offs that must result from various choices of k.

6.2 Managing the False Alarm Rate

In the application of MCUSUM and the MEWMA to the hospital respiratory chief complaint data, the MCUSUM produced a total of 35 signals and the MEWMA 36 signals over a two-year period. Given the on-going concerns with excessive false alarms in existing biosurveillance systems, it is worthwhile to assess how we got so many alarms when we set the thresholds to achieve an ATFS of 100 days.

In our application, a majority of the additional signals were redundant signals seemingly related to continuing outbreak conditions. In biosurveillance these types of redundant signals will tend to naturally arise as a result of the autocorrelation present in the data, both because outbreaks are persistent in time and because the adaptive regression residuals are not completely independent. For example, the first signal on June 12, 2003 was followed by nine more signals in the five weeks that followed it – one signal every 3-5 days (where the time between the signals occurred because the MCUSUM and MEWMA statistics were reset after each signal and thus it took a couple of days to the procedures to re-signal). Similarly, the first signal on December 20, 2003 was followed by 10 more signals, one every day or two right after the first signal subsequently spreading out to about a week apart later.

These types of continuing signals suggest a need for operator override capabilities in biosurveillance systems so that the operator can, for example, silence future signals *once an outbreak is identified*, much as one might silence a fire alarm once the fire trucks arrive. One can imagine, for example, a biosurveillance system designed so that once a procedure signals, is investigated, and an outbreak is confirmed – say it is a winter flu outbreak – the operator can denote the start of the outbreak in the system and turn off the signals until such time as the disease incidence rate returns to normal. For the application in Section 5, once these redundant signals are eliminated the number of seemingly true signals (two or three) plus the number of seemingly false first signals is reasonably close to what would be expected based on the chosen ATFS.

6.3 Biosurveillance for Bioterrorism Detection

The idea of silencing future signals once a natural disease outbreak occurs begs the question of the purpose of a biosurveillance system and whether during a winter flu outbreak the system should still be looking for a bioterrorism attack. Indeed, if the overriding goal is bioterrorism detection, then silencing the detection procedure during the flu season may be exactly the wrong thing to do. Furthermore, if that is the goal, then during a natural disease outbreak the biosurveillance systems will likely require "adjustment" to best detect a bioterrorism attack.

For example, consider the methods presented here or, for that matter, any method based on the adaptive regression with sliding baseline. If such a method is simply automatically implemented, as we did with the real data in Section 5, then an attack that occurs during a natural disease outbreak will be very difficult to detect. It will, of course, be difficult to detect anyway, but with a long sliding baseline the redundant signals that will occur as a result of the disease outbreak will be indistinguishable from the signals related to the attack.

Furthermore, the only way to (try to) identify an attack when an outbreak is occurring is to adjust the sliding baseline of the adaptive regression so that it only includes data from the outbreak period. If the outbreak is long enough that a reasonable new sliding baseline can be established, then it may be possible to look for signals that indicate a further departure from the outbreak. However, if that is the case, then the biosurveillance system will of necessity require a "human-in-the-loop" to identify the start of the flu season and to re-set the adaptive regression so that the sliding baseline only uses the flu outbreak data.

6.4 Can Biosurveillance be Automated?

While there is significant interest in the research and public health communities in devising automated methods of detection capable of sifting through masses of data to identify outbreaks and bioterrorism attacks, this research leads us to conclude that, at least for the foreseeable future, these types of detection algorithms alone will not be sufficient and human-in-the-loop strategies will be required for effective detection.

For example, the implementation of adaptive regression can be improved with the application of human judgement. As just described, if the purpose is bioterrorism detection, then the baseline should be adjusted – to the extent feasible – during natural disease outbreaks in order to make bioterrorism detection possible during such outbreaks. This can only be done by a human making an informed decision about when the outbreak started.

In a related vein, if the goal of a biosurveillance system is to detect departures from the natural background disease incidence, then the performance of the adaptive regression immediately after an outbreak will be improved by removing the outbreak data from the sliding baseline. That is, inclusion of the outbreak data could impair the accurate estimation of predicted values immediately following the outbreak. This again will require a human making an informed decision about both when an outbreak started and when it ended.

Finally, we note that simply allowing an adaptive regression to automatically proceed sequentially through biosurveillance data without removing outbreak data (as we allowed out of convenience in our example in Section 5) can result in false signals. This can occur when, for example, the sliding baseline consists mainly of the data during the period when an outbreak is subsiding, such that the slope of the regression is negative. Under this condition, if there is an abrupt return to the normal condition, then it is possible for the adaptive regression to under-predict, resulting in positive residuals and an alarm condition.

Simply put, to the adaptive regression methodology, a sudden increase from the background disease incidence pattern looks the same as a sudden leveling off from an outbreak returning back to a normal condition. And, while one might be tempted to not allow signals when, say, the slope of the adaptive regression is negative, this would also prevent the procedure from detecting outbreaks or attacks during periods when there are even slight improvements in disease incidence.

Of course, while this discussion has focused on the application of adaptive regression, the underlying issues are broader and not limited to just adaptive regression-based methods. This suggests that improvements in the practice of biosurveillance will continue to require advances in detection algorithms combined with development of detection algorithm practices and procedures as defined over time by experienced biosurveillance system operators and enhancements in biosurveillance system software capabilities.

Acknowledgements

The authors would like to express their appreciation to two anonymous reviewers for their helpful comments on the original draft, as well as to *Quality Engineering* for sponsoring this special issue. R. Fricker's research was partially supported by Office of Naval Research grant N0001407WR20172.

References

[1] Burkom, H.S., Murphy, S.P., and G. Shmueli (2006). "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine*, accepted (available at http://www3.interscience.wiley.com/cgi-bin/abstract/114131913/).

[2] Centers for Disease Control and Surveillance (2005). "Flu Activity: Reports & Surveillance Methods in the United States," www.cdc.gov/flu/weekly/fluactivity.htm, accessed December 14, 2005.

[3] Centers for Disease Control and Surveillance (2004). "Syndromic Surveillance: Reports from a National Conference, 2003," *Morbidity and Morality Weekly Report (Supplement)*, **53**, September 24, 2004.

[4] Chang, J.T. and R.D. Fricker, Jr. (1999). Detecting When a Monotonically Increasing Mean Has Crossed a Threshold, *Journal of Quality Technology*, **31**, 217–234.

[5] Crosier, R.B. (1988). Multivariate Generalizations of Cumulative Sum Quality Control Schemes, *Technometrics*, **30**, 291–303.

[6] Crowder, S.V. and Hamilton, M.D. (1992). An EWMA for Monitoring a Process Standard Deviation, *Journal of Quality Technology*, **24**, 12–21.

[7] Dunfee, D.A., and B.L. Hegler (2007). *Biological Terrorism Preparedness: Evaluating the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms*, Master's Thesis, Naval Postgraduate School, Monterey, CA.

[8] Espino, J.U., and M.M. Wagner (2001). The Accuracy of ICD-9–coded Chief Complaints for Detection of Acute Respiratory Illness, Proceedings AMIA Annual Symposium, 164-8.

[9] Follmann, D. (1996). A Simple Multivariate Test for One-Sided Alternatives, *Journal of the American Statistical Association*, **91**, 854–861.

[10] Fraker, S.E., Woodall, W.H., and S. Mousavi (2007). Performance Metrics for Surveillance Schemes, draft dated June 15, 2007.

[11] Fricker, R.D., Jr., Hegler, B.L., and D.A. Dunfee (2008). Comparing Syndromic Surveillance Detection Methods: EARS' Versus a CUSUM-based Methodology, *Statistics in Medicine*, DOI: 10.1002/sim.3197. Available on-line at http://www3.interscience.wiley.com/journal/96515927/issue.

[12] Fricker, R.D., Jr. (2007a). Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance*, **3**:1. Accessible on-line at www.isdsjournal.org.

[13] Fricker, R.D., Jr. (2007b). Syndromic Surveillance, *Encyclopedia of Quantitative Risk Assessment* (to appear).

[14] Fricker, R.D., Jr., and Rolka, H. (2006). Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, 4–13.

[15] Healy, J.D. (1987). A Note on Multivariate CUSUM Procedures, *Technometrics*, **29**, 409–412.

[16] Hu, C.X. and M.C. Knitt (2007). *A Comparative Analysis of Multivariate Statistical Detection Methods Applied to Syndromic Surveillance*, Master's Thesis, Naval Postgraduate School, Monterey, CA.

[17] Hutwagner, L., personal communication, December 12, 2006.

[18] Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr. (2008). A One-sided MEWMA Chart for Health Surveillance, *Quality and Reliability Engineering International* (to appear).

[19] Kleinman, K.P., Abrams, A., Mandl, K., and R. Platt (2005). Simulation for Assessing Statistical Methods of Biologic Terrorism Surveillance, *Morbidity and Mortality Weekly Report*, **54** (Supplement), Centers for Disease Control and Prevention. Accessed on-line at http://dacppages.pbwiki.com/f/mmwr2005.pdf on May 12, 2007.

[20] Kudô, A. (1963). A Multivariate Analogue of the One-Sided Test, *Biometrika*, **50**, 403–418.

[21] Lotze, T., Murphy, S.P., and G. Shmueli (2006). Preparing Biosurveillance Data for Classic Monitoring (draft), in submission to *Advances in Disease Surveillance*.

[22] Lowry, C.A., Woodall, W.H., Champ, C.W., and S.E. Rigdon (1992). A Multivariate Exponentially Weighted Moving Average Control Chart, *Technometrics*, **34**, 1, 46–53.

[23] Montgomery, D.C. (2001). *Introduction to Statistical Quality Control*, 4th edition, John Wiley & Sons, New York.

[24] Perlman, M.D. (1969). One-Sided Testing Problems in Multivariate Analysis, *The Annals of Mathematical Statistics*, **40**, 549–567.

[25] Pignatiello, J.J., Jr., and G.C. Runger (1990). Comparisons of Multivariate CUSUM Charts, *Journal of Quality Technology*, **3**, 173–186.

[26] Reis, B.Y., Pagano, M., and K.D. Mandl (2003). Using Temporal Context to Improve Biosurveillance, *Proceedings of the National Academy of Sciences*, **100**, 1961–1965.

[27] Roberts, S.W. (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, **1**, 239–250.

[28] Rogerson, P.A. and I. Yamada (2004). Monitoring Change in Spatial Pattens of Disease: Comparing Univariate and Multivariate Cumulative Sum Approaches, *Statistics in Medicine*, **23**, 2195–2214.

[29] Shmueli, G. (2006). Statistical Challenges in Modern Biosurveillance, in submission to *Technometrics* (draft dated September 18, 2006).

[30] Stoto, M.A. (2007). Public Health Surveillance, *Encyclopedia of Quantitative Risk Assessment* (to appear).

[31] Stoto, M.A., Fricker, Jr., R.D., Jain, A., Diamond, A., Davies-Cole, J.O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K., and C. Yuan (2006). Evaluating Statistical Methods for Syndromic Surveillance, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson, and D. Olwell, eds., New York: Springer.

[32] Testik, M.C., and Runger, G.C. (2006). Multivariate One-Sided Control Charts, *IIE Transactions*, **30**, 635–645.

[33] Woodall, W.H. (2006). The Use of Control Charts in Health-Care and Public-Health Surveillance, *Journal of Quality Technology*, **38**, 1–16.

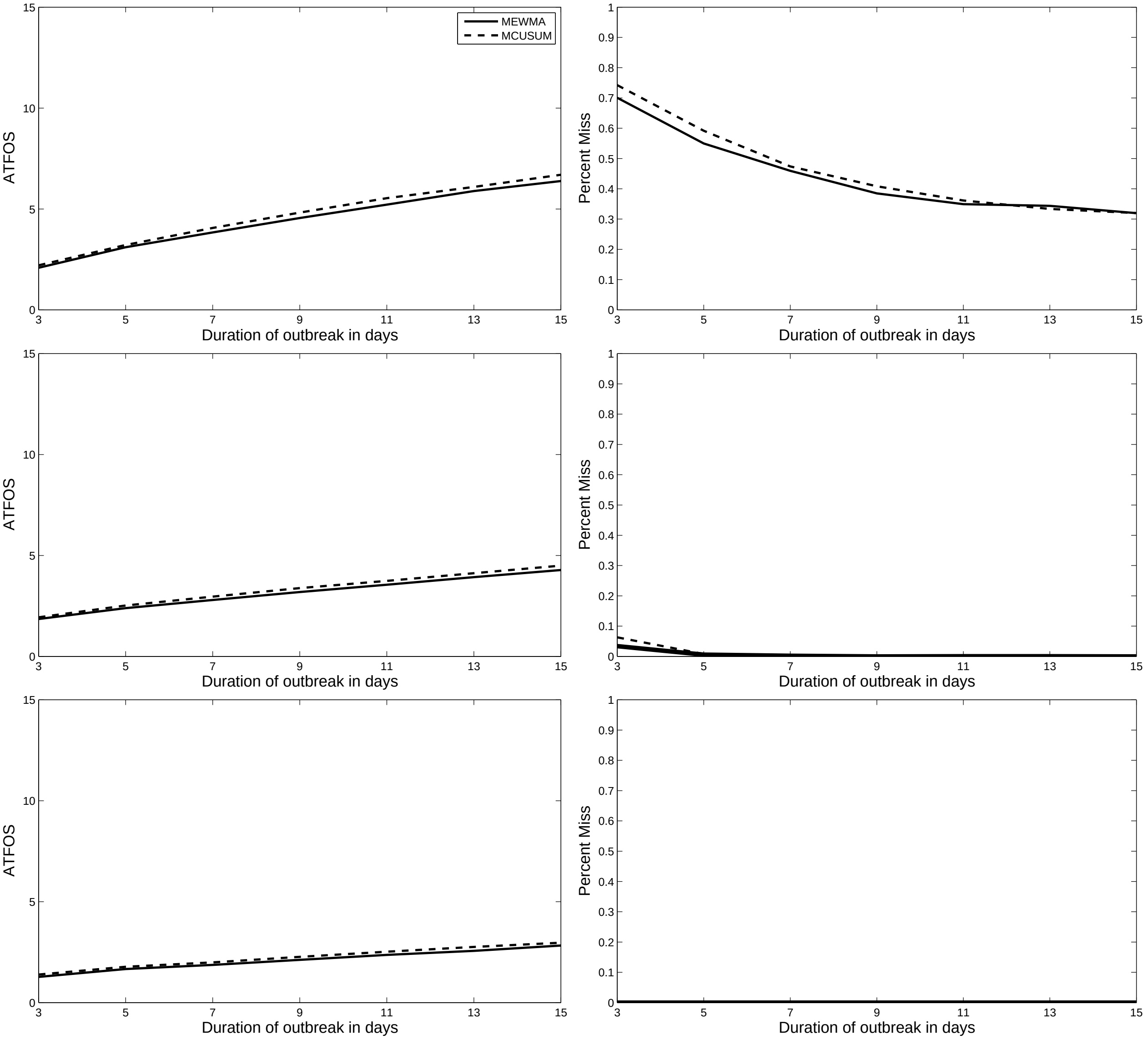

Figure 3: Performance of the MCUSUM and MEWMA under scenario 4 for three magnitudes of outbreaks – $M = 9$, $M = 22.5$, and $M = 45$, shown from top to bottom – versus various outbreak durations.

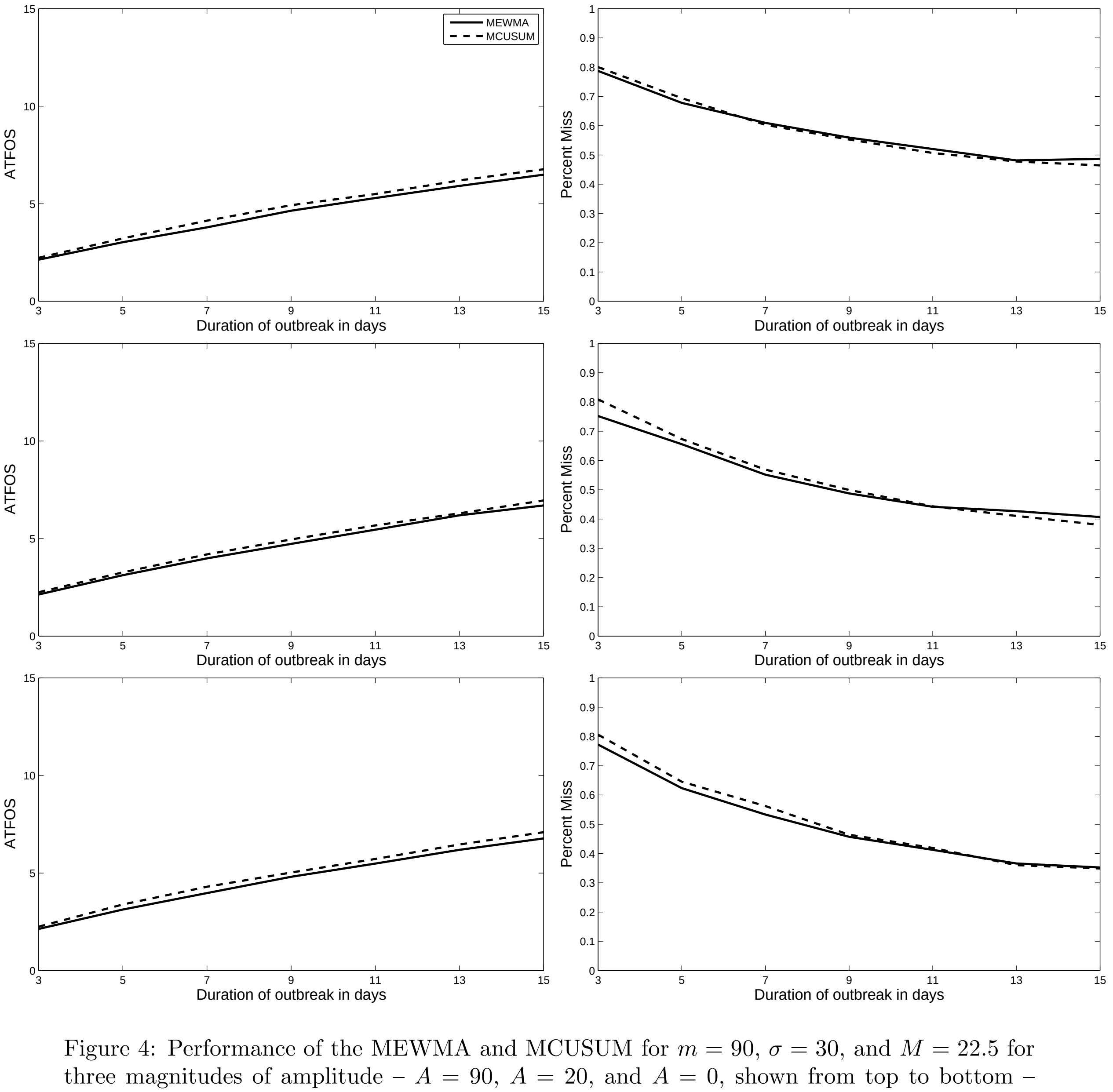

Figure 4: Performance of the MEWMA and MCUSUM for $m = 90$, $\sigma = 30$, and $M = 22.5$ for three magnitudes of amplitude – $A = 90$, $A = 20$, and $A = 0$, shown from top to bottom – versus various outbreak durations for $M = 22.5$.

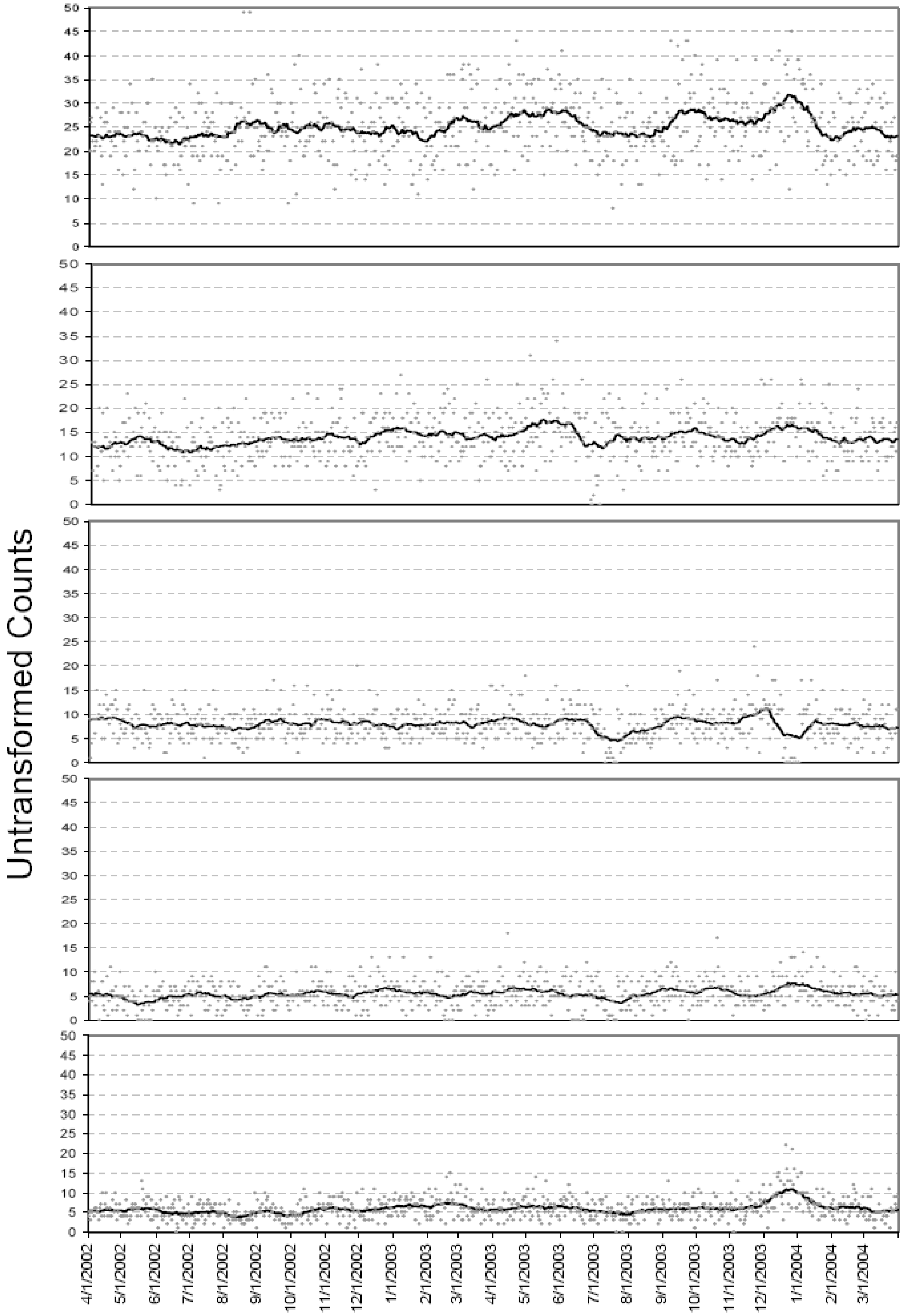

Figure 5: The data consisting of "chief complaint" respiratory counts for five hospitals with a smoothed mean line superimposed. The smoothed mean was calculated using a four week moving average from two weeks prior to two weeks after each day.

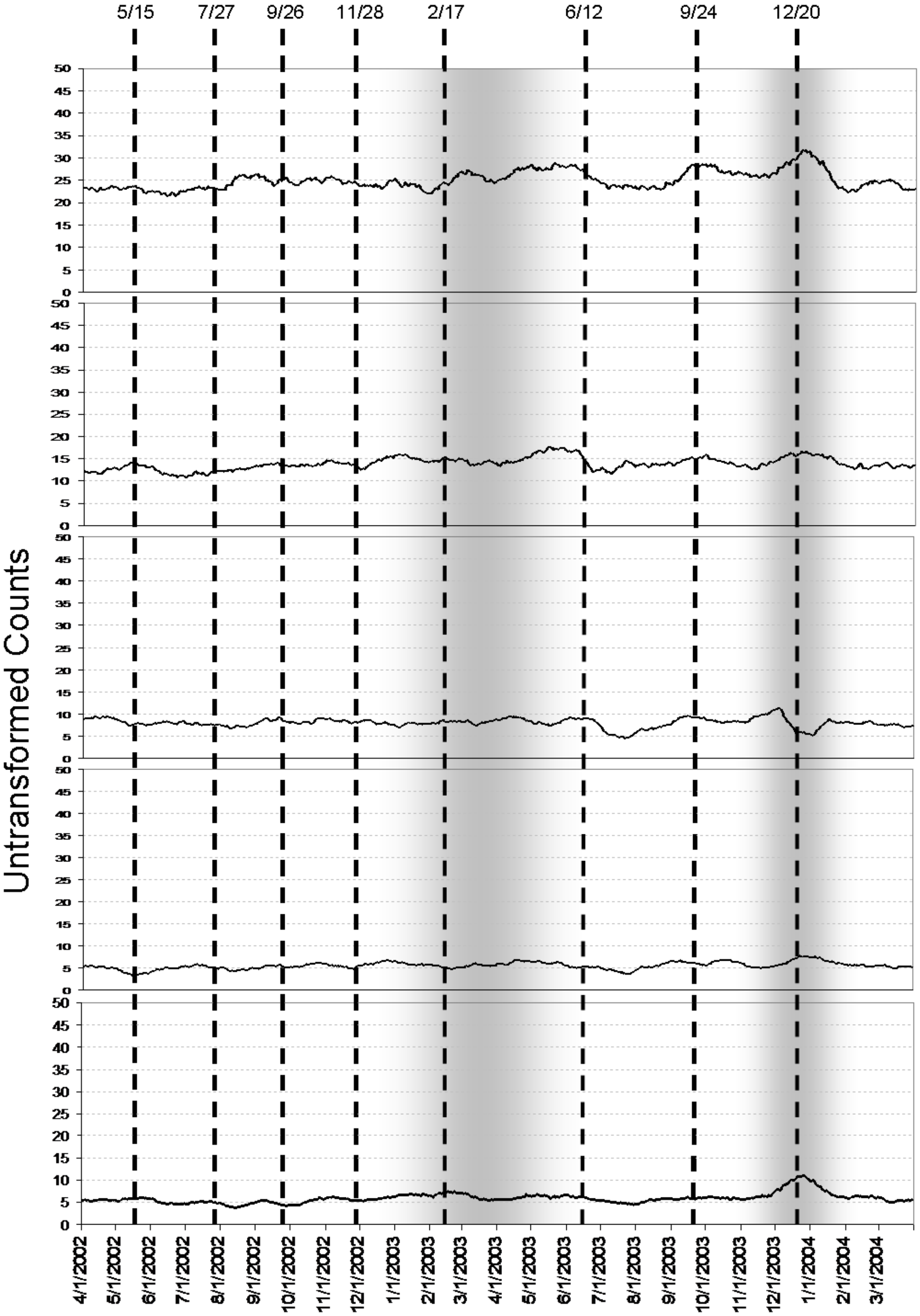

Figure 6: This plot shows when the modified MCUSUM and MEWMA first signaled when run on the future data. "First signaled" means that repeated signals within 60-days of the first signal are suppressed for plot clarity. The shaded areas indicate the flu seasons as indicated by the CDC's aggregate data on "percentage of visits for influenza-line illness reported by sentinel physicians."