Comparing Methods to Better Understand and Improve Biosurveillance Performance

“When I have fully decided that a result is worth getting I go ahead of it and make trial after trial until it comes.”

Thomas A. Edison
(1847-1931)

This chapter builds on Chapter 9, synthesizing everything from the first eight chapters (and Appendix B) to demonstrate how to compare the performance of early event detection (EED) methods using simulation. Assessing EED performance via simulation is critical because, as shown in the last chapter, comparisons using real data are limited in a number of ways.

The first limitation is that real data is opaque and often unknowable. For example, it’s virtually impossible to specify with certainty when a known outbreak starts and stops. It’s similarly impossible to state with certainty that data for a given period is free of outbreaks. The entire point of EED methods and tools like change point analysis is to find outbreaks, but failure to find an outbreak does not prove the absence of outbreaks.

The second limitation is that all real data is unique, both temporally and spatially. As a result, any analysis on one set of data is limited in terms of how generalizable the results might be for another time, or location, or under another set of circumstances. Given that EED methods will be broadly applied across many different populations, it is important that the performance of the methods is robust to such differences.

The third limitation is that outbreaks are (thankfully) scarce, as is the availability of real data sets (unfortunately). Hence, as shown in Section 9.2 of the last chapter, it’s very hard to reliably estimate ATFS, PSD, and CED, since for any given set of data the number of known/observable outbreaks will be small. The result is that comparisons between EED methods using real data will be inherently variable and thus clear and clean conclusions will be hard (at best) to reach.

Simulation overcomes many of these limitations. For example, with simulation the specifics of an outbreak can be precisely known and hence the performance of EED methods can be precisely quantified. In addition, many simulations can be conducted over a wide variety of conditions, from which
the robustness (or lack thereof) of various EED methods can be judged. Furthermore, simulated data can easily be shared, facilitating collaboration and allowing for replication of research results.

The major criticism of simulation is that simulated data cannot capture all of the complexity inherent in real data. This is both a strength and weakness of simulation. It is a weakness if simulated data fails to mimic the important and salient features of real data, where it is important to note that not all real world data complexity is necessarily relevant. However, it is also a strength in the sense that the appropriate abstraction often facilitates the identification of the specific conditions under which a method is more or less effective.1

This chapter contains two examples of using simulation to compare EED methods. The first compares the univariate CUSUM to the EARS methods. It’s drawn from joint research with Ben Hegler and Andy Dunfee (Fricker et al., 2008a). The second compares the MCUSUM and MEWMA methods and is drawn from joint research with Matt Knitt and Cecilia Hu (Fricker et al., 2008b).

CHAPTER OBJECTIVE

The goal of this chapter is to demonstrate the use of simulation for comparing the performance of EED methods. The results presented herein are meant to be illustrative, both of one simulation approach and of the utility of simulation for comparing EED performance. While the results are informative with respect to the performance of the particular EED methods compared, definitive conclusions should follow only after these results have been replicated and additional comparisons under other conditions have been made.

---

1 This point brings the text full-circle back to the quotes in the Preface on page VII.
MATHEMATICAL NOTATION

\begin{align*}
A & \quad \text{Amplitude of the sinusoid simulating seasonal variation} \\
c & \quad \text{Mean level of disease in UTM1 simulation} \\
C_t & \quad \text{CUSUM statistic for day } t \\
D & \quad \text{Outbreak duration} \\
\delta, \delta & \quad \text{Day-of-the-week effect} \\
\mathbb{E}(Y) & \quad \text{Expected value of random variable } Y \\
\gamma, \gamma & \quad \text{Seasonal deviation from annual mean} \\
h & \quad \text{EED method threshold} \\
I & \quad \text{Identity matrix} \\
k & \quad \text{MCUSUM/CUSUM parameter} \\
\lambda & \quad \text{MEWMA smoothing parameter} \\
M & \quad \text{Outbreak magnitude} \\
\mu & \quad \text{Vector of mean disease in MTM1 simulation} \\
n & \quad \text{Length of sliding baseline} \\
\mathcal{N}(\mu, \sigma^2) & \quad \text{Univariate normal distribution with mean } \mu \text{ and variance } \sigma^2 \\
\mathcal{N}_5(\mathbf{0}, \sigma^2 \mathbf{I}) & \quad \text{Five dimensional multivariate normal distribution with} \\
& \quad \text{mean vector } \mathbf{0} \text{ and covariance matrix } \sigma^2 \mathbf{I} \\
o, o & \quad \text{Mean outbreak level} \\
\sigma_Z & \quad \text{Standard deviation of random variable } Z \\
\tau & \quad \text{Random day of outbreak start} \\
\theta, \theta & \quad \text{Within-season deviation} \\
U(a, b) & \quad \text{Uniform distribution over the range } a \text{ to } b \\
y_t & \quad \text{Observation } y \text{ on day } t \\
Y, \mathbf{Y} & \quad \text{Random variable/vector representing a count before it is observed} \\
Z, \mathbf{Z} & \quad \text{Random noise component in data simulations}
\end{align*}
10.1 Performance Comparisons: A Univariate Example

This section illustrates how to conduct simulation comparisons between some of the univariate EED methods presented in Chapter 7. In so doing, it brings together many concepts discussed in previous chapters, including adaptive regression from Chapter 5, as well as methods for simulating biosurveillance data from Appendix B. In particular, this section compares the EARS C1, C2, and C3 methods to the CUSUM method applied to the forecast errors of an adaptive regression model.

10.1.1 Simulating Biosurveillance Data

In order to compare the EARS and CUSUM EED methods, background disease incidence was simulated and then various types of simulated bioterrorism attacks/natural disease outbreaks were overlaid. The background disease incidence data was simulated using the Univariate Temporal Method #1 (UTM1) of Appendix B as the sum of a mean disease incidence, a seasonal sinusoidal cycle, a systematic day-of-the-week effect, and random noise. Outbreaks, when they occurred, were incorporated as an additive term.

That is, a daily observation $Y_t$ was simulated as

$$Y_t = \max(0, \left[ c + \gamma_t + \delta_t + o_t + Z_t \right]), \quad t = 1, 2, 3, \ldots$$  \hspace{1cm} (10.1)

where

- $c$ is a constant level of disease incidence;
- $\gamma$ is the seasonal deviation;
- $\delta$ is the day-of-the-week effect;
- $Z$ is the random noise around the systematic component, $c + \gamma_t + \delta_t$;
- $o_t$ is the mean outbreak level which, when an outbreak is occurring, increases the disease incidence level as described below; and,
- $\lceil x \rceil$ is the ceiling function, which rounds $x$ up to the next largest integer.

The seasonal effect is calculated as $\gamma_t = A[\sin(2\pi t/365)]$, where $A$ is the maximum deviation from $c$ with $t = 1$ corresponding to October 1st on a 365 day per year calendar. For the random noise, the simulations use $Z \sim N(\mu, \sigma^2)$ when $c$ is large and $Z \sim LN(\mu, \sigma^2)$ when $c$ is small.

The day-of-the-week effect is the systematic deviation from $c + \gamma_t$, where $\delta_t = \delta_{t+7}$ for all $t$. It is defined in terms of $\sigma$, a parameter of $Z$: $\delta = -0.5\sigma$ on Sunday, $\delta = 0.1\sigma$ on Monday, $\delta = 0.2\sigma$ on Tuesday, $\delta = 0.3\sigma$ on Wednesday, $\delta = 0.4\sigma$ on Thursday, $\delta = 0$ on Friday, and $\delta = -0.3\sigma$ on Saturday.

Table 10.1 specifies parameter values for Equation 10.1 for 12 scenarios designed to span a range of possible underlying disease incidence patterns. Scenarios 1-6 are large-count scenarios and Scenarios 7-12 are low-count scenarios. The parameters were selected to generate synthetic data that mimic
disease incidence patterns similar to selected data sets published by the CDC (www.bt.cdc.gov/surveillance/ears/datasets.asp).

In particular, $c = 90$, $A = 80$, $\mu = 0$ and $\sigma = 30$ or $\sigma = 10$ in Equation 10.1 result in disease incidence patterns similar to EARS data set S08. Setting $c = 90$, $A = 20$, $\mu = 0$ 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. For Scenarios 7-12, combinations of the values in Table 10.1 result in disease incidence patterns similar to S03, S04, S15, and S34. Figure 10.1 shows one set of simulated data for Scenario 1.

For the low count scenarios, data set S04 is characteristic of hospital-level respiratory or influenza-like illness (ILI) chief complaint counts, S03 of hospital-level rash chief complaint counts, and S34 of hospital-level neurological chief complaint counts. For the high count scenarios, S08 is characteristic of state-level aggregate respiratory or influenza-like illness (ILI) chief complaint counts, S45 of state-level aggregate gastrointestinal chief complaint counts, and S15 of state-level neurological chief complaint counts.

Outbreaks were incorporated into Equation 10.1 as an additive term $o_t$ representing the mean outbreak level. As with the simulated data itself, the outbreak is an idealized form that could be parameterized simply. The parameters are the peak magnitude $M$, the outbreak duration $D$, and a random start day $\tau$. Outbreaks increase linearly up to $M$ and then linearly back down to zero:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>c</th>
<th>A</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>80</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>80</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>20</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>20</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>6</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>6</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 10.1. Parameters for Equation 10.1 that define 12 background disease incidence scenarios.
Fig. 10.1. One year of simulated Scenario 1 data.

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

The performance of the EED methods were assessed for outbreaks of various magnitudes and durations.

- Scenarios 1-6 used three magnitudes – small, medium, and large – defined as a fraction of the constant disease incidence \( c \): \( M = 0.1c \), \( M = 0.25c \), and \( M = 0.5c \), respectively, where \( c = 90 \) from Table 10.1.

- Scenarios 7-12 used four magnitudes – very small, small, medium, and large – defined as a fraction of the mean plus three standard deviations of the lognormally distributed random variable \( Z \) from Equation 10.1.
  - Large outbreaks: \( M = \mathbb{E}(Z) + 3\sigma_Z \);
  - Medium outbreaks: \( M = 0.5[\mathbb{E}(Z) + 3\sigma_Z] \);
  - Small outbreaks: \( M = 0.25[\mathbb{E}(Z) + 3\sigma_Z] \);
  - Very small outbreaks \( M = 0.1[\mathbb{E}(Z) + 3\sigma_Z] \).

For Scenarios 7-12, \( \mathbb{E}(Z) = \exp(\mu + \sigma^2/2) \)

and

\[ \sigma_Z^2 = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}, \]

where \( \mu \) and \( \sigma \) are specified in Table 10.1 for the various scenarios.

For all the scenarios considered, durations ranged from short to long: \( D = 3, 5, \ldots, 15 \) days. The simulations were conducted in MatLab 7.1.0.246.
using the \texttt{randn} function to generate random normal variates and \texttt{lognrnd} to generate lognormal random variates.

### 10.1.2 Determining the CUSUM Parameters

In order to use the CUSUM EED method on the forecast errors of an adaptive regression, a number of choices need to be made. First, the form of the adaptive regression must be chosen (linear, quadratic, etc.) and the length of the sliding baseline \((n)\) must be set. For the CUSUM, the reference value \(k\) must be chosen. How these choices were made are briefly discussed here. See Fricker \textit{et al.} (2008a) for additional details.

All other factors being equal, adaptive 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 potentially a better prediction, often in biosurveillance the amount of available data is limited or the older data is 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.

To assess this tradeoff, for each of the 12 scenarios in Table 10.1, Dunfee & Hegler (2007) conducted detailed preliminary simulations to find the form of the adaptive regression that forecast best. Figure 10.2 is an example of how the form of the adaptive regression and size of the “optimal” sliding baseline for Scenario 2 \((c = 90, A = 80, \mu = 0, \sigma = 10, \text{with and without day-of-the-week effects})\) were determined. The optimal \(n\) was chosen by visual inspection with the criteria that the \(n\) be as small as possible but also as close to achieving the best forecast in terms of the minimum average squared residual.

Figure 10.2 shows that the linear model achieved almost the same minimum average squared residual as the quadratic model but with a smaller \(n\). As described in Dunfee & Hegler (2007), this occurred consistently for all of the scenarios leading them to choose a linear adaptive regression model in all of their evaluations. For the linear model, across all the scenarios without day-of-the-week effects, the optimal \(n\) values ranged from 15 to 40 days. For the scenarios with day-of-the-week effects, the optimal \(n\) values increased, with the largest being around 56 days – the size recommended by Burkom \textit{et al.} (2006).

As discussed in Section 7.3, a common setting for the CUSUM reference value \(k\) is one-half the distance between the mean of the normal disease incidence and the level of incidence it is important to quickly detect. Here the CUSUM is not being applied to the raw data, but rather to the standardized residuals from an adaptive regression, so the change in mean can be expressed in terms of some number of standard deviations increase in the standardized residuals. If a one standard deviation increase in the residuals is important
Fig. 10.2. 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 2. On the left are the results when there is no day effect in the data and on the right the data has day effects. From this, a linear model form for the adaptive regression, with an “optimal n” of about 30 days for no day effects and about 40 days with day effects, was chosen.

to detect, because there is also some effect of the length of the baseline (see Fricker et al. 2008a for details) for those CUSUMs with larger sliding baselines (e.g., 30 ≤ n ≤ 60) set k = 1/2; for the CUSUMs using a seven-day sliding baseline (designed to match the baseline used in the EARS methods), set k = 0.65.

10.1.3 Comparison Methodology

The usual metrics are used to compare performance: ATFS, PSD, and CED. For each EED method and scenario in Table 10.1, the threshold for each method required to achieve an ATFS of 100 days was determined empirically. Once the thresholds were set, the methods were then compared across all the scenarios specified in Table 10.1 for all the outbreak types just described.

The purpose of setting the thresholds to achieve equal time between false alarms was to ensure a fair comparison between the methods. That is, it is always possible to improve a method’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 it is then possible to make an objective judgement about which method is best at detecting a particular type of outbreak.

Across all the scenarios in Table 10.1, the CUSUM thresholds ranged from $h = 2.9$ to $h = 4.2$, including all combinations of $c, A, \mu, \sigma$ and with and without day-of-the-week effects. For the EARS methods, across all of the
scenarios in Table 10.1 the thresholds for C1: $2.7 \leq h \leq 8.2$; for C2: $2.6 \leq h \leq 7.4$; and for C3: $3.0 \leq h \leq 18.2$.

Having set the thresholds to achieve equal ATFS performance, the CED and PSD were calculated as follows. For each iteration $i$, the methods 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 method signalled during this time it was re-set and restarted, just as it would be in a real application. This allowed the CUSUM 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 method signaled at time $t_i$ within the duration of the outbreak, the time to first outbreak signal was recorded as $t_i - 100$ and the steady-state CED was estimated as $\sum_{i=1}^{s}(t_i - 100)/s$ for the $s$ iterations that signaled within the outbreak duration. The PSD was calculated as the number of iterations for which the method signaled during the outbreak divided by the total number of iterations run.

10.1.4 Results

Figure 10.3 in many ways summarizes the results of all the evaluations conducted by Dunfee & Hegler (2007). In it, the plots on the left side show the conditional expected delay (CED) versus various outbreak durations ($D$) for Scenario 2 starting with a small outbreak at the top ($M = 9$), a medium outbreak in the middle ($M = 22.5$), and a large outbreak at the bottom ($M = 45$). The plots on the right side show the probability of successful detection (PSD) of an outbreak versus outbreak duration. Each plot gives the results for six methods, the C1, C2, and C3, as well as three CUSUMs using various sliding baseline lengths: 7, 15 (the “optimal” for Scenario 2), and 56 days.

What Figure 10.3 shows is that the C1, C2, and C3 methods do not perform as well as the CUSUM methods with the larger sliding baselines. Focusing for a moment just on the C1, C2, and C3 methods, note that the C1 and C2 methods perform somewhat similarly, with the C1 generally having a slightly lower CED compared to the C2 but at the expense of having a slightly lower PSD as well. However, when comparing the C1 and C2 to the CUSUMs, note that they all have similar CED performance but the CUSUMs with longer sliding baselines have much higher PSDs. This difference in performance is evident for all the outbreak magnitudes, but is most striking with the larger magnitude outbreaks. For example, in the middle row of plots, the C1 and C2 CED can be up to a day or so shorter than the longer sliding baseline CUSUMs, but they only catch between about 20-30 percent of the outbreaks while the 56-day sliding baseline CUSUM catches nearly 80 percent of the outbreaks of the longest duration. For this scenario, it is clear that the CUSUM with a 56-day sliding baseline is the preferred method.
Fig. 10.3. CED and PSD performance of the methods for Scenario 2 for three magnitudes of outbreaks – $M = 9$, $M = 22.5$, and $M = 45$, shown from top to bottom – versus various outbreak durations.
A note about the CED plots is in order for those used to looking at graphs of average run lengths in the statistical process control literature. Such readers may be surprised that the CED curves increase as outbreak duration increases. Remember in the biosurveillance problem that the time to first outbreak signal is constrained to the interval \([1, D]\). That is, the earliest a “true signal” can occur is on the first day of the outbreak and the latest is on the last day of the outbreak \((D)\). Thus, for \(D = 3\), the CED is constrained to be between 1 and 3 and, as shown in the plot, is about 2 for all the methods. On the other hand, for \(D = 15\) the CED can be much larger and, in fact, falls anywhere from about 3.5 days to about 6.5 days for the various methods.

Also in Figure 10.3, the C1 and the CUSUM with a 7-day sliding baseline suffer from being contaminated by the outbreak data in the largest magnitude outbreak scenarios. That is, in the lower right plot the PSD for these two methods actually decreases for longer duration outbreaks (as eventually does the C2 and C3, as well as the CUSUM with a 15-day sliding baseline ever so slightly). If these methods fail to detect the outbreak early on, they begin to incorporate the outbreak data into their calculations (either the moving average for the C1 or the adaptive regression predictions for the CUSUM), making it increasingly more difficult to distinguish the outbreak from the normal background disease incidence. In comparison, the two-day lag in the C2 method seems to be sufficient to eliminate much of this problem for that method (and the C3 which is a function of the C2 statistics), at least for the outbreaks of smaller duration (i.e., \(3 \leq D \leq 7\) or so).

Figure 10.4 shows the results for Scenario 7. What is immediately striking between Figures 10.3 and 10.4 is the overall similarity of the CUSUM performance results. The CUSUMs with the longer sliding baselines are clearly the best performing methods (where note that the “optimal” sliding baseline in this scenario was 30 days, compared to 15 days for Scenario 2). Within the EARS methods, the C1 method has the lowest CED but also misses more outbreaks than the C2, and neither perform as well as the CUSUMs with longer sliding baseline. In particular, while the C1 does have the shortest CED of all the methods, it completely misses from 85 to 90 percent of all the outbreaks. In comparison, in the bottom plots with the larger magnitude outbreaks for example, the CUSUMs using either a 30- or 56-day sliding baseline catch virtually all of the outbreaks with a 2-day CED for a 3-day outbreak duration up to a 4-day CED for a 15-day outbreak duration.

10.1.5 Discussion

In these comparisons the CUSUM applied to the residuals of adaptive regressions perform better than the EARS methods. This conclusion follows because the EARS methods frequently failed to catch a majority of the outbreaks across a wide variety of background disease incident patterns (large and small daily counts; large, medium, small, and no seasonal cycles; large
Fig. 10.4. Performance of the methods for Scenario 7 for three magnitudes of outbreaks – $M = 4$, $M = 8$, and $M = 16$, shown from top to bottom – versus various outbreak durations.
and small random daily fluctuations; with and without day-of-the-week effects) and a wide variety of outbreak magnitudes and durations. In fact, the EARS methods generally caught less than 30 percent of the outbreaks except in the largest outbreak cases. In contrast, the CUSUM methods, particularly with the 8-week sliding baseline, performed much better.

These conclusions are based on extensive comparisons of the methods using simulated biosurveillance data that was designed to mimic the major features of a wide cross-section of biosurveillance data. However, as noted earlier, the simulations were purposely idealized depictions that assumed a sinusoidal shape for the annual background variation, linearly increasing and decreasing outbreaks, and particular error term distributions. In addition, these analyses used a fixed ATFS of 100 days, which is a reasonable false alarm rate for a biosurveillance system, and a particular choice for the the CUSUM reference interval parameter $k$. However, confirmation of whether these results hold more generally under other conditions requires further research.

Given the performance of the CUSUM methods, particularly those using longer sliding baselines, one might be tempted to simply attribute the success to the additional information being used in the adaptive regressions. This is certainly part of the reason, but some additional preliminary simulations seem to indicate that is not the complete answer. Specifically, it seems that Shewhart and Shewhart-based methods may be less well suited for the biosurveillance problem in which outbreaks do not occur instantaneously and are transient.

For example, Figure 10.5 compares the performance of a CUSUM method and a Shewhart method, both applied to the residuals from an adaptive regression with a 56-day sliding baseline. Each day, the Shewhart method compares the standardized residual from the adaptive regression for that day to a threshold (chosen so that the ATFS is 100 days; the same as with the CUSUM). In Figure 10.5, the top plots compare the methods for a fixed outbreak magnitude ($M = 22.5$) for varying outbreak durations. The bottom plots compare the methods for a fixed outbreak duration ($D = 7$ days) for various outbreak magnitudes. The background disease incidence was generated via Scenario 2.

The top plots show that the CED is roughly equal between the two methods, with the Shewhart seeming to have a slight advantage for outbreaks of short duration and the CUSUM for outbreaks with long durations. This is consistent with the literature on the performance of these two methods in industrial statistical process control applications. However, the upper right plot shows that the Shewhart is much poorer at actually catching outbreaks than the CUSUM.

In the bottom two plots, the magnitude is varied rather than the outbreak duration. That is, the duration is fixed at $D = 7$ days and then the performance of the two methods is compared as $M$ varied from 10 to 70. In terms of CED, the Shewhart does slightly better than the CUSUM for smaller outbreaks and slightly worse for larger outbreaks. But, once again,
it does significantly poorer in terms of PSD. From this, a conjecture is that
the poorer performance of the EARS methods may be due both to the additional data used in the CUSUMs with the longer sliding baselines and to the Shewhart-like design of the C1 and C2 methods.

In summary, the CUSUM applied to residuals from an appropriately employed adaptive regression model with an 8-week sliding baseline outperformed the EARS methods in all the scenarios Dunfee & Hegler (2007) evaluated. These scenarios were chosen to mimic the major features of biosurveillance.

Fig. 10.5. Performance comparison of the CUSUM versus Shewhart method applied to the residuals of an adaptive regression with an 8-week sliding baseline. The top plots compare the methods for a fixed outbreak magnitude \(M = 22.5\) for varying outbreak durations. The bottom plots compare the methods for a fixed outbreak duration \(D = 7\) days for various outbreak magnitudes. The background disease incidence was generated via Scenario 2.
data over a wide variety of conditions. For standard biosurveillance systems using the EARS methods, this suggests biosurveillance systems may benefit from replacing the EARS methods with CUSUM methods and from setting the CUSUM thresholds appropriately to minimize the false alarm burden as much as is appropriate.

Of course, the EARS methods were originally designed for a drop-in surveillance system with little or no baseline data available. In these situations the use of an 8-week sliding baseline may be impossible, at least upon initiation of the drop-in system. However, the simulations showed that a CUSUM with a 7-day sliding baseline performed about the same as the EARS methods, and as the length of the sliding baseline increased the performance of the CUSUM quickly improved. This suggests a strategy for drop-in surveillance systems of starting with a CUSUM with a 7-day sliding baseline and, as time progresses and more data accumulates, allowing the baseline to increase until such time as enough data is accumulated so that baseline can be allowed to slide.

10.2 Performance Comparisons: A Multivariate Example

As with the previous section, this section illustrates how to conduct simulation comparisons between the multivariate EED methods presented in Chapter 8, specifically the MCUSUM and the MEWMA. In so doing, it brings together many concepts discussed in previous chapters, such as simulating biosurveillance data from Appendix B and adaptive regression from Chapter 5, as well as the multivariate methods from Chapter 8.

10.2.1 Simulating Multivariate Biosurveillance Data

To begin, using scenarios similar to those in Table 10.1, the Multivariate Temporal Method #1 (see page 383) is used to simulate background disease incidence data for a biosurveillance system monitoring the GI syndrome from five hospitals. As with the univariate comparisons in the last section, the goal is to assess how the multivariate methods perform over a range of outbreak magnitudes and durations.

To simulate the background data for Scenario 4, the parameters used in Equation B.8 are:

- \( \mu = \{90, 90, 90, 90, 90\} \);
- \( \gamma_t = \{\gamma_{t1}, \gamma_{t2}, \gamma_{t3}, \gamma_{t4}, \gamma_{t5}\} \);
- \( \theta_t = \{0, 0, 0, 0, 0\} \); and,
- \( Z_t \sim N_5(0, \sigma^2I) \), with \( \sigma = 10 \).

In this simulation day-of-the-week effects (\( \delta \)) were omitted and \( \gamma_{tj} = 20[\sin(2\pi(t + \tau_j)/365)] \) with \( \tau_j \sim U(-30, 30) \) for \( j = 1, \ldots, 5 \). More realistic
Comparing Methods to Better Understand and Improve Biosurveillance Performance

Simulations would likely use $\mu$ and $\gamma$ levels that vary between the hospitals, and they could incorporate day-of-the-week effects and other model terms, but the purpose here is to illustrate how to conduct a performance comparison under some relatively simple conditions. Given this background disease incidence, outbreaks ($o$) were simulated using Equation B.5 with $M = 5, 10, 20$ and $D = 3, 4, \ldots, 14, 15$ days.

10.2.2 Determining the MEWMA and MCUSUM Parameters

As discussed with the univariate methods, to use the MCUSUM and MEWMA EED methods on the forecast errors of an adaptive regression, a number of choices need to be made. Just like before, the form of the adaptive regression must be chosen (linear, quadratic, etc.) and the length of the sliding baseline ($n$) must be set. For the MCUSUM, the reference value $k$ must be chosen and for the MEWMA the smoothing parameter $\lambda$ must be chosen. How these choices were made are briefly discussed here. See Fricker et al. (2008b) for additional details.

In the standard statistical process control setting, it is well known that the standard univariate EWMA can be designed (through the appropriate choice of $\lambda$) to perform very similarly to the CUSUM. Furthermore, the way to choose the reference value $k$ in the univariate CUSUM follows directly from its derivation, but the choice for $k$ in Crosier’s MCUSUM is not clear.

Very little research into these issues has been conducted in a biosurveillance context. Indeed, the literature only contains one attempt to address these questions, where Fricker et al. (2008b) found that (for $p = 4$) the MCUSUM detection method with $k = 0.74$ performed very similarly to the MEWMA detection method with $\lambda = 0.2$.

Figure 10.6 shows the results they obtained where, to simplify their analysis, the evaluation was conducted using the standard statistical process control assumptions of iid observations and a sustained jump change in the mean. Then, comparing how well the MCUSUM detected the various sustained mean shifts for a four dimensional standard multivariate normal, Figure 10.6 shows that $k = 0.74$ gives the closest performance to the MEWMA with $\lambda = 0.2$.

That is, Figure 10.6 shows the percent change in the ATFS for the MCUSUM for various $k$ compared to the MEWMA. The plot shows 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. The MCUSUM with $k = 0.74$ has 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.
Thus, for the purposes of these simulation comparisons, for the MEWMA $\lambda = 0.2$ was used and for the MCUSUM $k = 0.7$ was used. In terms of setting thresholds, quantile-quantile plots of the residuals confirmed that they are reasonably normally distributed. Thus, the tables in Appendix C can be used to set the thresholds so that each detection method will achieve an ATFS of 100 days. From the tables for $p = 5$, the MCUSUM detection method with $k = 0.8$ for $h = 5$ gives ATFS=103.1. For the MEWMA detection method with $\lambda = 0.2$ for $h = 12$ gives ATFS=98.

Finally, the form of the adaptive regression and the optimal length of the sliding baseline was evaluated, much as was done by Dunfee & Hegler (2007) with very similar results. The linear form had smaller sliding baselines that achieved almost the same minimum average squared residuals. And, for the linear model, across all the scenarios, the optimal $n$s 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 $n$s were larger with the largest being around 56 days.

10.2.3 Comparison Methodology

The multivariate comparisons in this section were conducted the same way as the univariate methods in the last section.
Figures 10.7 and 10.8 summarize the main findings: the MEWMA and MCUSUM performed virtually identically, both in terms of CED and PSD, across all the scenario and outbreak combinations evaluated. Though the lines deviate slightly in Figures 10.7 and 10.8, the differences are not statistically significant. Thus, it seems that, just like for the univariate methods in the classical SPC situation, with an appropriate choice of parameters, the MCUSUM and MEWMA EED methods can be made to perform virtually identically for biosurveillance.

Specifically, the plots in Figure 10.7 show 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 scenarios. For example, Figure 10.8 shows the results for Scenarios 1, 3, and 5 for an outbreak of medium magnitude. See Knitt & Hu (2007) for plots for all of the scenarios and types of outbreaks.

Figure 10.7 demonstrates how the procedures perform for the various types of outbreaks. For example, the CED 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 CED 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.

Furthermore, Figure 10.8 demonstrates that the adaptive regression with sliding baseline methodology does very well at removing the systematic component, at least for this synthetic biosurveillance data. Here 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 CED, there is no visible difference between the three plots in Figure 10.8. 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.

### 10.3 Discussion & Summary

Since the performance of most proposed methods is demonstrated on data that is not publicly available, it is very difficult and often impossible to compare
Fig. 10.7. 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.
Fig. 10.8. 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 $M = 22.5$. 
the performance of the various detection methods across the biosurveillance literature. This is often driven by the public health community’s desire to see methods demonstrated on real data. Yet, precisely because the data is real, there is a lack of general availability of such data to the research community due to confidentiality and privacy concerns.

One solution is to make real data more widely available. Shmueli & Burkrom (2010) say, “Currently syndromic data are only available to researchers affiliated with a particular biosurveillance system or research group, for reasons of data confidentiality and non-disclosure agreements. This a major obstacle in the way of scientific progress in both temporal and spatio-temporal biosurveillance, and hopefully some data will be made available to academic researchers.”

However, seemingly unrecognized in this discussion is the fact that any real data is simply one realization of a stochastic process. Focusing only on a particular stream of data one fails to recognize and account for the full randomness of the underlying phenomenon, which is a process involving the interaction of very complicated population and disease transmission dynamics.

In addition, even if some real data is made available, it will provide little to no information about what outbreaks look like, particularly those associated with bioterrorism related events. The challenge, as Rolka et al. (2007) have said, “...is to develop improved methods for evaluating detection algorithms in light of the fact that we have little data about outbreaks of many potential diseases that are of concern.”

Simulation is an alternative. Of course, it is very difficult to (stochastically) characterize, and thus simulate, all the detailed features characteristic of the normal or baseline state of disease incidence, as well as the various outbreak conditions. However, one could also make similar statements about industrial quality control problems. Yet that field, over time, has come to use various data abstraction conventions that facilitate simulation and, as a result, allow comparisons between methods and across the literature. As Rolka et al. (2005) said, “Reliance on the use of Monte Carlo simulation in the field of Statistics is well known. It has been this author’s experience that the technique is undervalued in the field of Public Health because it has previously not been required.”

As this chapter has demonstrated, simulation can be useful in evaluating the performance of EED methods, particularly their relative performance. In addition, simulation is useful for:

- evaluating methods across many scenarios;
- eliminating unnecessary/distracting real world complexities;
- allowing clean and clear comparisons of methods; and,
- making it easier to get at generalizable conclusions/results.