Detecting disease outbreak regions using multiple data streams

Sesha Dassanayake[1] 0000-0002-8662-5436 and Joshua P. French[2] 0000-0002-9708-3353)

1 *Department of Mathematics and Computer Science, Loyola University New Orleans, New Orleans, LA;*

2 *Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO*

Corresponding author: Sesha Dassanayake, Department of Mathematics and Computer Science, Loyola University New Orleans, New Orleans, LA – 70118, USA; skdassan@loyno.edu

Detecting disease outbreak regions using multiple data streams

Abstract

A novel approach for biosurveillance using multiple data streams is presented. The proposed method is computationally simple, has rapid detection ability, and produces few false alarms. The proposed algorithm is applied to three popular statistical process control (SPC) charts: the Shewhart Chart, the EWMA, and the CUSUM. The proposed method collects disease counts from multiple data streams, computes charting statistics, and then compares these to empirical in-control distributions generated using bootstrap methods to decide whether to signal an alarm. As bootstrap methods are used, no assumption is made about the in-control distributions corresponding to a specific parametric distribution – an assumption that is common with most conventional SPC methods. The proposed method relies on p-values and controls the false discovery rate; which distinguishes it from traditional SPC methods. The relatively low false alarm rate is a highlight of the proposed method, as higher false alarm rates are a common problem with conventional SPC charts. Through extensive simulations, the EWMA and CUSUM methods are shown to have superior performance over the widely-used Shewhart charts, with the EWMA having a slight advantage over the CUSUM. The proposed method is applied to the 2011 E.coli outbreak in Germany.

Keywords: Biosurveillance; CUSUM control chart; EWMA control chart; false discovery rate; Shewhart control chart; statistical methods for disease surveillance

1. Introduction

Disease surveillance systems use data to signal the existence of a potential outbreak based on statistical anomalies in the observed data. Early detection is critical in disease surveillance, as the main objective is to inform the public health authorities as early as possible so that harmful consequences from an outbreak can be reduced. Traditional disease surveillance systems confirmed outbreak occurrence retrospectively. Confirmed laboratory results were collected, analyzed and reported to decision makers after the outbreak occurrence. As Shmueli and Burkom (19) explain "… in most situations they (public health data) are collected, delivered, and analyzed days, weeks, or even months after.” As a result, with this kind of analysis, it is not possible for the decision makers to take any preventive measures. In the late 1990s, research focus shifted from these retrospective methods to biosurveillance, which is “… the practice of monitoring data to detect, investigate, and respond to disease” (19). Biosurveillance is performed prospectively to detect an outbreak as early as possible.

During the past two decades, many methods have been proposed for biosurveillance (19). Some of these methods use statistical process control charts (SPC) that originated in industrial process control. Woodall (22) pointed out that SPC methods have a long history of applications to problems in biosurveillance. Control charts are the main tools used in statistical process control to monitor quality characteristics of industrial processes. The Shewhart Chart, the exponentially weighted moving average chart (EWMA), and the cumulative sum chart (CUSUM), are popular statistical process control charts. Of the three charts, the Shewhart chart has been widely used for disease surveillance. For example, the Shewhart chart has been used to monitor anesthesia related adverse effects (7) and risk-adjusted mortality rates in patients admitted to hospitals for myocardial infarction (4). The CUSUM chart has also been frequently applied in biosurveillance (8). Elbert and Burkom (6) discussed an adaptation of the EWMA control chart for biosurveillance. Fricker (8) pointed out that “the EWMA, although popular in industrial SPC, is less commonly used in biosurveillance.”

We propose a method for detecting disease outbreaks that improves on weaknesses of traditional SPC methods in several ways: (i) our approach uses empirical p-values based on the bootstrap method, as suggested by Li et al. (12), allowing the research to quantify strength of evidence for an outbreak (or more accurately, one can quantify how incompatible the data are with a specified no outbreak model), (ii) because we must monitor multiple attribute variables, such as disease counts from multiple geographic regions, we address the multiple comparisons problem by controlling the false discovery rate (FDR), as suggested by Li and Tsung (11) , in order to improve the power of our method in comparison to methods that control the familywise error rate (FWER), which can be unnecessarily restrictive, (iii) the proposed method uses the FDR-controlling procedure proposed by Storey and Tibshirani (21) for FDR control, instead of the more conservative FDR control methods proposed by Benjamini and Hochberg (1).

This paper is organized in the following manner. Section 2 gives preliminary information about the standard Shewhart, EWMA, and CUSUM methods. Section 3 provides details about the proposed methodology. A comparison of the proposed method with standard methods is also provided. Section 4 describes the results of a simulation study conducted to compare the performance of this methodology in combination with the Shewhart, EWMA, and CUSUM methods. Initial conclusions based on these extensive simulation studies are also provided in this section. Section 5 describes an application of the proposed method to the 2011 *E.coli* outbreak in Germany. Finally, Section 6 explains the pros and cons of the proposed method, along with directions for future work.

2. Preliminaries

The proposed algorithm was applied to three popular control charts: (i) Shewhart Chart (ii) EWMA, and (iii) CUSUM. We briefly describe each of these methods in more detail.

2.1. Shewhart chart

Shewhart (18) invented the control chart to determine when the mean of an industrial process has changed. A Shewhart chart is defined by three characteristics: the centerline, the lower control limit (LCL), and the upper control limit (UCL). The process is declared to be out-of-control when the value of the observed process falls outside the interval [LCL, UCL].

Let the mean of the process that we are interested in monitoring be μ and the standard deviation be σ. The centerline of the Shewhart chart is μ. The LCL and UCL are defined in Montgomery (14) as

$$LCL = \mu - L\sigma,$$

$$UCL = \mu + L\sigma,$$

where L defines how far the upper and lower control limits are from the centerline.

2.2. Exponentially weighted moving average chart (EWMA)

Roberts (17) introduced the EWMA control chart for independent, identically distributed (iid) normal random variables. Borror *et al.* (2) extended the EWMA to Poisson data.

Assume $Y_1, Y_2, \dots, Y_n$ are iid Poisson counts with mean μ observed at times $t = 1, 2, \dots, n$. When the monitoring process is in control, we assume $\mu = \mu_0$. Borror *et al.* (2) proposed monitoring the process behaviour at time t using the test statistic

$$E_t = \lambda Y_t + (1 - \lambda) E_{t-1},$$

where λ is a weighting constant such that $0 < \lambda \leq 1$, and the starting value E_0 is set to the in-control mean or the target count rate μ_0.

Montgomery (14) recommends using the following UCL and LCL for a Poisson EWMA control chart with centerline μ_0:

$$UCL = \mu_0 + A_U \sqrt{\frac{\lambda \mu_0}{2-\lambda}[1-(1-\lambda)^{2t}]}, \tag{1}$$

$$LCL = \mu_0 - A_U \sqrt{\frac{\lambda \mu_0}{2-\lambda}[1-(1-\lambda)^{2t}]}, \tag{2}$$

where A_U and $A_L$are upper and lower control limit factors and the expression in square roots is the in-control process standard deviation. In many applications, the limits are chosen to be symmetric around zero, so that $-A_L = A_U$ (14). After determining the in-control mean μ_0, usually from historical data, and specifying λ, standard tables such as the tables provided by Borror *et al.* (2) can be used to select a control limit factor A that would give a desired in-control average run length ARL_0.

The ARL_0 is related to the tolerable false alarm rate. The centerline specifies where the process characteristic should be when the process is in control. When the process is in control, nearly all of the sample points fall between the upper and the lower control limits. Sometimes, even when the process is in control, a sample point can still be plotted outside of the control limits: this is a false alarm and is similar to Type I error in classical hypothesis testing. ARL_0 is defined as the average time between two such false alarms.

In contrast to an industrial process control setting, where out-of-control processes are likely to be under the LCL or above the UCL, in a disease surveillance setting, we are only interested in identifying mean disease counts above what is typically seen during a non-outbreak period. In that context, Shu *et al.* (20), proposed the one-sided EMWA statistic to detect upward shifts from the mean:

$$E_t = \max[\mu_0, \lambda Y_t + (1-\lambda)E_{t-1}]. \tag{3}$$

In this one-sided context, an alarm is signalled when the statistic in Equation (3) is above the UCL in Equation (1).

2.3. The cumulative sum chart (CUSUM)

The CUSUM control chart was proposed by Page (15) for iid normal responses. Later, Lucas (13) extended the method to iid Poisson responses. Consider the iid Poisson responses $Y_1, Y_2, \dots, Y_n$ with mean λ taken at times $t = 1, 2, \dots, n$. The control chart is designed to signal an alarm when the process mean shifts from an in-control mean of $\lambda = \lambda_0$ to an out-of control mean of $\lambda = \lambda_1$. The monitoring CUSUM statistic is defined as

$$C_t = \max(0, C_{t-1} + Y_t - k), \tag{4}$$

where Y_t is the observed count at time *t*, C_t is the CUSUM statistic at time *t*, and

$$k = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}, \tag{5}$$

is a constant value defined to minimize the time to detect a mean shift from the in-control mean of λ_0 to the out-of-control mean of λ_1.

The starting value of the CUSUM chart C_0 is often set to $C_0 = 0$. An alarm is signaled when $C_t > h$, where the threshold h is pre-determined. In fact, h is a function of both k and ARL_0. After the desired ARL_0 is specified and the constant k is determined using λ_0 and $\lambda_1$values, the threshold h can be determined using Monte Carlo simulations or statistical tables.

2.4. Multivariate statistical process control charts using p-values and controlling FDR

Conventional statistical process control charts use control limits to signal alarms. An alarm is signalled when the charting statistic exceeds pre-determined control limits. Li *et al.* (12)

proposed a method using p-values instead of the traditional control limits to signal alarms. They proposed two methods for computing p-values in a control chart setting: we use their second method, based on bootstrapping. Suppose we want to compute a p-value for a statistic at monitoring time t. Then t observations are sampled with replacement from the in-control observations and treated as a time series of *t* in-control observations; the charting statistic is computed over time for these *t* observations to get the charting statistic at time *t*. This process is repeated for B bootstrap samples, resulting in B bootstrap statistics at time t. Let $\{S_t^{*(1)} \dots, S_t^{*(B)}\}$ denote the set of charting statistics computed at time t for each of the B bootstrap samples. The distribution of these statistics represents the bootstrap distribution or the typical distribution of the charting statistic at time t when the process is in control. In classical hypothesis testing, this is referred to as the null distribution. Next, during the monitoring period, the usual statistic is computed for each measurement taken from the industrial process, with S_t denoting the charting statistic at time t. The observed test statistics are compared with the bootstrap distribution to calculate a p-value using the formula

$$\frac{1 + \sum_{j=1}^{B} I\left(S_t^{*(j)} \geq S_t\right)}{B + 1}.$$

The method proposed by Li *et al.* (12) is so general that it can be used with most commonly used control charts. We used this approach with all three of the control charts – Shewhart, EWMA, and CUSUM.

The method proposed by Li *et al.* (12) was for a single control chart. However, in order to monitor multiple geographic regions simultaneously, we need to use multiple control charts. When multiple control charts are used, we run into the multiple testing problem that inflates the probability of a false alarm. Li and Tsung (11) used false discovery rate (FDR) to address the multiple testing problem using conventional SPC charts with traditional control limits. We also use FDR to address the multiple testing problem: however, we use p-values instead of traditional control limits.

Benjamini and Hochberg (1) popularized the use of FDR to handle the multiple testing problem. The false discovery rate is defined as the expected proportion of false discoveries among all discoveries. Formally, FDR is defined as

$$\mathrm{FDR} = \mathrm{E}[V/R],$$

where V is the number of false alarms and R the number of total alarms. Several methods have been proposed to control FDR, such as the widely used Benjamini Hochberg procedure (1). We used a method more recently proposed by Storey and Tibshirani (21) to address the multiple testing problem in genomewide studies when the tests may be correlated.

3. Methods

3.1. The proposed method

Consider a study area partitioned into m regions. The disease count in region i at time t is Y_{it}, and the set of disease counts across all regions at time t is $\boldsymbol{Y}_t = \{Y_{1t}, \dots, Y_{mt}\}$. We will assume that this disease count represents the number of cases observed in the past week, which is a common reporting time period.

We detail the steps of the proposed method:

We start by specifying the tolerable level of false discoveries α. For example, if $\alpha = 0.05$, then on average we expect 5 false alarms for every 100 alarms.

Next, we identify a non-outbreak period from historical data. For example, for weekly data the non-outbreak period could be the most recent year without an outbreak. Thus, we can potentially update the baseline period every year in order to account for changing factors such as population change.

Suppose we want to perform surveillance over a period of L time steps. Let $\{\boldsymbol{Y}_1^0, \dots, \boldsymbol{Y}_s^0\}$ denote the collection of disease counts across the m regions for s times during the most recent non-outbreak period. We want to create B bootstrap samples of L observations using the data from the non-outbreak period. Let $\mathbf{Y}^{*(j)}$ denote the jth bootstrap sample. A single bootstrap sample $\mathbf{Y}^{*(j)} = \{\boldsymbol{Y}_1^{*(j)}, \dots, \boldsymbol{Y}_L^{*(j)}\}$ is obtained by sampling L observations with replacement from $\{\boldsymbol{Y}_1^0, \dots, \boldsymbol{Y}_s^0\}$, with $\boldsymbol{Y}_b^{*(j)} = \{Y_{1b}^{*(j)}, \dots, Y_{mb}^{*(j)}\}$ denoting the bth bootstrap observation.

During the surveillance period the following steps are performed.

1. Create B bootstrap samples, $\{\mathbf{Y}^{*(1)}, \dots, \mathbf{Y}^{*(B)}\}$ using most recent non-outbreak data, as discussed above.
2. For each bootstrap sample, the charting statistic is computed across the L monitoring times for each of the m regions. The charting statistic for bootstrap sample j for region i at time step t is denoted $S_{it}^{*(j)}$.
3. The charting statistics at time t across all L times are computed for each region, with S_{it} denoting the charting statistic of region i for time t.
4. The p-value of the ith region at time step t is computed as
$$p_{it} = \frac{1 + \sum_{j=1}^{B} I\left(S_{it}^{*(j)} \geq S_{it}\right)}{B + 1}.$$
At each time step, there will be a vector of associated p-values $\boldsymbol{p}_t = (p_{1t}, \dots, p_{mt})$. Figure 1 provides a flow chart describing this basic process.

Finally, to address the multiple testing problem, we use the ST procedure for FDR control. FDR control relies on knowing the number of tests that are being considered. Since we are trying to control the FDR across m regions but a potentially unknown number of time steps, we control the FDR with respect to each individual time step, but not necessarily across the entire data stream of a single region.

The final output of the algorithm is a set of m alarm decisions from which we can identify the regions that need attention.

3.2. Comparison with existing methods

We now describe several advantages of the proposed method for disease surveillance over conventional methods.

First, p-values are more flexible for testing than the traditional control limits. A p-value is a fundamental statistical quantity used in a variety of disciplines, whereas control limits are primarily limited to industrial process control. Testing via control limits results in a binary decision of either an alarm or no alarm. In contrast, a p-value quantifies the strength of incompatibility of the data with the null hypothesis. Thus, even when there is no alarm, a p-value can still be used to assess whether there may be an embryonic outbreak in its earliest stages.

Second, the FDR is a more useful error criterion for disease surveillance than ARL_0. The ARL_0 is defined as the expected run length time between false alarms. In industrial process control, a process is calibrated to be in-control when it starts. When the charting statistic exceeds the control limits, an alarm is signalled and the process is stopped, recalibrated, and started again. In disease surveillance, we cannot stop an outbreak after an alarm, making the control of ARL_0 of little value in that context.

Additionally, ARL_0 is directly related to the Type I error. In a disease surveillance setting Type I error would indicate the probability of an alarm when there is no outbreak. However, a non-detected outbreak is likely to be of greater concern than incorrectly concluding there is an outbreak. Consequently, it is more natural to control the expected proportion of false alarms, which is what FDR measures. The additional benefit of controlling the FDR instead of ARL_0 is an expected increase in testing power. Li and Tsung (11) note that in conventional SPC charts, “the control limits are determined by fixing the overall in-control average run length. However, the power of such Bonferroni-Type control charts is rather low, when the number of charts is large." By choosing to control the FDR, it is possible to use more powerful and up-to-date multiple testing procedures, such as the ST method, in contrast to methods designed to control the FWER.

Third, the proposed method makes modest distributional assumptions, increasing its applicability to observed data. Conventional SPC methods typically rely on theoretically defined alarm thresholds based on parametric assumptions. However, the proposed method does not assume the disease counts follow a specific distribution, as the empirical non-outbreak distribution is generated using bootstrap methods. Buckeridge *et al*. (3) recommend using empirical methods as opposed to relying on traditional SPC distributional assumptions.

Fourth, in-control parameters need to be regularly updated in a disease surveillance setting. When monitoring disease rates over geographic regions for extended periods of time, one needs to consider various changes that occur over time that affect the typical mean disease count observed during a non-outbreak period. These changes may include medical coding changes related to disease diagnosis, new treatments such as vaccines, population shifts, changes in data participation or information systems, and other factors. Similarly, in-control parameter estimates should be re-estimated in an industrial process control context when the underlying process is thought to have changed. All three of the methods employed in what follows – the Shewhart, EWMA, and CUSUM methods - need to be regularly updated with the "in-control mean" or the expected mean disease count during a non-outbreak period. By utilizing the most recent non-outbreak time period to obtain the in-control distribution, our method allows these parameters to change gradually over time without additional user intervention.

There has been some similar research that enables FDR control in charting statistics. Lee et al (10) assume that data come from a iid normal distribution, and define a statistic that follows the standard normal distribution asymptotically when the null hypothesis is true (i.e. when there is no outbreak). Then, p-values are calculated using the standard normal distribution.

Most conventional SPC methods rely on this assumption that data come from an iid normal distribution. However, Buckeridge et al. (3) point out that “Public health surveillance data tend to violate assumptions of (conventional) SPC methods”. The proposed method makes minimal distributional assumptions. Bootstrap samples are collected from a non-outbreak period (an in-control period) from which the corresponding bootstrap statistics are calculated. Empirical in control distributions are generated using these bootstrap statistics from which p-values are calculated. Buckeridge et al. (3) recommend using empirical methods as opposed to relying on traditional SPC distributional assumptions. The proposed method is more suitable for public health surveillance data that violate traditional distributional assumptions.

Dassanayake and French (5) presented a spatio-temporal method using the CUSUM statistic. A spatio-temporal statistic was calculated pooling counts of neighbouring regions to signal alarms. The proposed method is a purely temporal method (using three popular SPC charts - the EWMA, the CUSUM, and the Shewhart chart), which considerably simplifies the computational burden only using the disease count data at each time step.

3.3. Summary

The first step of the proposed method is to set a tolerable FDR level $\boldsymbol{\alpha}$. Next, using historical data, a non-outbreak period is identified. The non-outbreak period is regularly updated for each region to account for relevant factors changing over time. E.g., the weekly disease counts from the most recent year without an outbreak are used as the in-control period. The proposed method is applied using the output from the desired SPC chart (Shewhart, EWMA, or CUSUM). Afterwards, bootstrap samples are taken from each region during the non-outbreak period. For these bootstrap samples, the appropriate SPC chart is calculated. An empirical in-control (i.e. non-outbreak) distribution for each region is developed using these bootstrap statistics. Then at each time step *t* after the non-outbreak period, disease counts are collected from each region and the relevant SPC method is applied to the counts for that region. These statistics are compared with the corresponding in-control distributions for each region to determine p-values. As there is a p-value from each region at each time step, we get a vector of p-values. Finally, for each time step, the ST procedure is used to address the multiplicity problem encountered when testing multiple p-values simultaneously.

4. Simulation experiment

4.1. Experiment setup

Disease counts were simulated in 36 contiguous regions arranged in a 6×6 grid. The time period for the simulation was 100 time points, where each time point represented a "day." In each of the 36 regions, independent Poisson counts were simulated. The first half of the simulation (days 1-50) represented the no outbreak period and the second half (days 51-100) represented the outbreak period. For the no outbreak period from days 1-50, independent Poisson counts were simulated in each of the 36 regions with a constant mean of $\lambda_{0i} = 4,\ i = 1, \dots, 36$. However, for the outbreak period from days 51-100, independent Poisson counts were simulated with out-of-control means (denoted by $\lambda_{1i},\ i = 1, \dots, 36$) peaking at the central regions and gradually thinning towards the perimeter regions, as illustrated in Figure 2. Note that in all regions along the perimeter, no outbreak was simulated. Also note that, spatial contiguity of the data streams in Figure 2 has nothing to do with the analysis but helps with visualization of a stylized scenario.

[Figure 2 near here]

No outbreak was simulated in the 20 perimeter regions, regions 1, 2, 3, 4, 5, 6, 7, 12, 13, 18, 19, 24, 25, 30, 31, 32, 33, 34, 35, and 36. Corner regions in the inner 4 x 4 grid (shaded in medium dark grey), specifically, regions 8, 11, 26, and 29 get a 1 standard deviation shift in the mean. The remaining 8 perimeter regions in the inner 4 x 4 grid (shaded in light grey) – regions 9, 10, 14, 17, 20, 23, 27, and 28 – get a 2 standard deviation shift in the mean. Finally, the four central regions – regions 15, 16, 21, and 22 – gets a 3 standard deviation shift in the mean.

In this simulation, the proposed method was implemented using three popular statistical process control charts: (i) the Shewhart, (ii) the EWMA, and (iii) the CUSUM control charts.

First, the proposed method was applied using 36 Shewhart charts – one chart for each region. $B = 10{,}000$ bootstrap samples were randomly selected from the non-outbreak period to construct an empirical in-control distribution (null distributions) for each region. At each time step, p-values for each region were calculated by comparing the simulated disease counts at each time step with the corresponding empirical distribution for each region. The FDR level for each test was set to 0.05.

Second, the proposed method was implemented using 36 EWMA statistics – one EWMA statistic for each region. The weighting constant λ was set to 0.20 (Montgomery (14) recommends selecting a λ value where $0.05 < \lambda \leq 0.25$.) An in-control mean of $\lambda_0 = 4$ was used as the starting value E_0 for each chart. An empirical in-control distribution for each region was generated using $B = 10{,}000$ bootstrap samples randomly selected from in-control data for each region. The FDR level for each test was set to 0.05.

Third, the proposed method was applied using 36 CUSUM statistics – one statistic for each region, like the implementation of the EWMA statistics. Each CUSUM statistic was designed with an in-control Poisson mean (λ_0) of 4 and an out-of-control Poisson mean (λ_1) of 6, 8, 10, depending on the region. Thus, each CUSUM statistic was designed to detect a change of one standard deviation increase in the mean. Similar to the EWMA method, an empirical in-control distribution for each of the 36 regions was generated using $B = 10{,}000$ bootstrap samples randomly selected from the non-outbreak period. The FDR level was set to 0.05, as before.

4.2. Simulation results

Figure 3 shows the results for a single simulation over a 100-day period for region 26. Recall that for region 26, the in-control mean is 4 from days 1-50 and the out-of-control mean is 6 from days 51-100. Region 26 is one of the four regions – 8, 11, 26, 29 – to receive the smallest shift in the mean of 1.0 standard deviation. Figure 3 (a) shows the simulated disease counts for the surveillance period from days 1-100. Figure 3 (b) shows the standardized statistics for all three methods: the statistics were standardized by dividing each statistic by its maximum value within the surveillance period from days 1-100. Note that the standardized EWMA statistic (red) has a higher reference line compared to the other two standardized statistics –Shewhart (blue) and CUSUM (dark green) – both of which have a reference line of zero. The reason is that in calculating the EWMA statistic, we plot the maximum of the two numbers, the in-control mean for the region or the current EWMA statistic at time *t*, following equation (3) from Section 2.2 for the EWMA computation. However, with the CUSUM statistic, we plot the maximum of the two numbers, zero or the current CUSUM statistic at time *t*, using equation (4) from Section 2.3 for the CUSUM computation. Therefore, it is possible for the CUSUM statistic to return to zero, unlike the EWMA. In addition, the Shewhart statistic shows the standardized observed counts during the surveillance period. Therefore, the Shewhart statistic does not return to zero as all simulated counts during the surveillance period are positive values. Figure 3(c) shows the alarms signalled by each method. The Shewhart statistic (blue) signals the first alarm on day 72 – 21 days after the onset of the outbreak on day 51. However, the CUSUM statistic (dark green) signals the first alarm 9 days earlier than the Shewhart statistic on day 63. The quickest to signal alarms is the EWMA statistic which signals the first alarm on day 61. Furthermore, the Shewhart statistic signals alarms sporadically – only four alarms on days 72, 85, 93, and 98 - during the outbreak period from days 51 -100. However, the CUSUM statistic is more persistent in signalling alarms during the outbreak period signalling a total of 38 alarms during the outbreak period from days 51-100. The EWMA method also signals alarms rather persistently over the outbreak period – signalling a total of 33 alarms– even for a moderate 1.0 standard deviation shift in the mean in this region.

[Figure 3 near here]

Figure 4 shows the results for region 22, one of the four center regions in Figure 2 with the largest shift in the mean. In region 22, the in-control mean changes from 4 to 10 at the onset of the outbreak on day 51. The Shewhart method (blue) starts to signal 2 days after the start of the outbreak, on day 53. Note that the alarms are more frequent during the outbreak period for region 22 (with a 3 standard deviation increase in the mean) compared to region 26 (with a 1 standard deviation increase in the mean). The CUSUM method (dark green) also detects the outbreak on day 53. Of all three methods, the EWMA method is the quickest to signal the outbreak on day 52. Also note that all three methods signal alarms more frequently in region 22 than in region 26, as the shift in the mean is much larger: the EWMA method signals continuously after detecting the outbreak on day 52; the CUSUM also signals continuously after detecting outbreak on day 53; Shewhart method signals more persistently in region 22 than in region 26.

[Figure 4 near here]

In order to verify that all three methods were controlling FDR at 0.05 level, 100 independent simulations following the method outlined in Section 3.2 were carried out using the three statistics. The empirical FDR was computed at each time step for the three independent simulations and the mean of these statistics was computed. The mean empirical FDR was 0.044 for the EWMA with a 95% confidence interval of 0.040 to 0.047; CUSUM had a mean FDR of 0.049 with a 95% confidence interval of 0.046 to 0.052; the Shewhart had a mean FDR of 0.044 with a 95% confidence interval of 0.042 to 0.047. Thus, for all three methods the empirical FDR was controlled at each time step across regions. The mean FDR values for both the EWMA and the Shewhart were comparable; compared to both EWMA and Shewhart statistics, the mean FDR was slightly higher for the CUSUM as it takes a little longer for the CUSUM to decrease after signalling an alarm.

We also computed statistics related to the power of each method. The empirical power is the percentage of the time that an outbreak is detected across all regions at times that an outbreak occurred. Shewhart charts had an empirical power of 28.57% with a 95% confidence interval of

28.02% to 29.04%; CUSUM had 96.19% mean power with a confidence interval of 96.00% to 96.41%; EWMA had a 90.23% mean power with a confidence interval of 89.93% to 90.62%.

4.3. Comparing performance of the three methods

In evaluating the performance of the three surveillance methods, certain measures should be adopted. Frisen and Sonesson (9) highlight that "good properties (of a disease surveillance system) are quick detection and few false alarms." They list (i) conditional expected delay (CED) and (ii) probability of a false alarm (PFA) as two common measures for evaluating the speed of detection and rate of false alarms, respectively. Conditional expected delay (CED) is the average delay time until an alarm when the change occurs at time point τ, and defined as

$$CED(\tau) = \mathrm{E}[t_A - \tau | t_A \geq \tau].$$

In the event of an outbreak, CED is the expected number of time periods from the beginning of the outbreak to the first alarm. The other measure, probability of a false alarm is defined as,

$$PFA = \mathrm{P}(t_A < \tau).$$

In other words, this is the probability of an alarm before the actual outbreak.

In calculating the CED and PFA measures, data sets were generated with a range of change points (5, 10, 15, …, 95). For example, for the first change point $\tau = 5$, 100 data sets were simulated with no outbreaks from days 1-4 having an in-control Poisson mean (λ_0) of 4; for the outbreak period from days 5 – 100, Poisson data were simulated with out-of-control means (λ_1) specified in Figure 2. Then for each data set, the delay from the start of the outbreak on day number 5 was calculated. For instance, if the first signal after the onset of the outbreak on day 5 occurred, let's say, on day 7 for region 1, then the delay in detection for that region was calculated to be 2 days. Similarly, delays in detection were calculated for each of the 36 regions in all 100 data sets. Next, the delays were averaged for each region. In other words, for each region there were 100 detection delay measures calculated from the 100 data sets. These detection delays were averaged for each region to determine the CED values for the change point when $\tau = 5$. Likewise, for the other change points ($\boldsymbol{\tau} = 10, 15, 20, \ldots, 95$), the CED values were computed for each region. In order to make accurate CED calculations for change points near the end of the 100-day surveillance period, the simulation time period was extended to 150 days, with outbreaks extending from their start until time 150. By using multiple change points for the start of the outbreak, we are able to assess the impact of having more or less null data when applying the methodology: in other words, we assessed the impact of different warm-up periods.

Figure 5 shows the CED values versus change points for regions 9, 11, and 16 – three regions that experience the outbreak during the outbreak period. Since there are regions that do not experience the outbreak during the outbreak period such as regions 1, 2, and 3, only a selection of regions experiencing the outbreak were illustrated as CEDs cannot be calculated for regions not experiencing outbreaks. Recall that region 11 received a 1 standard deviation shift in the mean, region 9 a 2 standard deviation shift in the mean, and region 16 a 3 standard deviation shift in the mean. One-way ANOVA tests for all 16 regions that experience the outbreak in the inner 4 x 4 region were conducted using the CED values for all three statistics. It is important to observe that in all three regions – regardless of the magnitude of the shift in the in-control mean, the EWMA has the lowest CED values, implying speedier detection. In the four regions that receive the highest shift (3 standard deviation) in the mean, namely regions, 15, 16, 21, and 22, the Shewhart statistic which uses the current disease count was faster than the CUSUM which cumulates current and past observations. In all other regions in the inner 4 x 4 region, the

CUSUM was faster than the Shewhart. The key finding was that the EWMA was the quickest to detect an outbreak in all regions.

A careful study of Figure 5 shows how the speed of detection for the three methods improves with the magnitude of the increase in the mean. Recall that the magnitude of the increase in mean gradually increases from region 11, to region 9, to region 16, with region 16 having the largest increase in the mean. With the increase in the magnitude of the mean shift, the CED values steadily decrease for all three methods. For the Shewhart method the CED values for regions 11 are between 9.5 – 18.8. With the increase in the mean shift for region 9, the CED values for Shewhart statistic decrease to values between 3.2 – 4.4. With the largest increase in the mean, for region 16, the CED values further decrease to values between 0.8 – 1.5. This pattern can be observed for both the CUSUM and the EWMA methods: with higher shifts in the mean, the CED values systematically decrease to lower values. All three methods are generally speedier in detecting larger shifts in the mean.

[Figure 5 near here]

Figure 6 shows the PFA values for the three methods versus change points. Similar to the simulation for the CED, in calculating PFA values, 100 data sets were simulated for the same change points ($\tau = 5, 10, 15, \ldots, 95$). For example, for the change point $\tau = 5$, in-control data were simulated for days 1-4 and out-of-control data for days 5-100. For each simulation, PFA values were calculated as the proportion of false alarms during the no outbreak period from days 1-4. For all 100 simulations with change point $\tau = 5$, PFA values were calculated for each region and averaged, as before. The same calculation was carried out for all other change points

($\tau = 10, 15, 20, \ldots, 95$). Figure 6 shows PFA values versus change points for the same three regions: region 9, 11, and 16. A careful analysis using one-way ANOVA (at $\alpha = 0.05$ level of significance) for all 16 regions that experience the outbreak in the inner 4 x 4 region revealed that Shewhart had the highest PFA for all 16 regions. There was no statistically significant difference in PFA between the EWMA and CUSUM (except for regions 11, 20, and 26 where the CUSUM has a slightly higher PFA). Figure 6 also reveals the extremely low occurrence of false alarms for all three statistics using the proposed algorithm. Note the scale of the y-axis - the highest PFA value for all three plots is below 0.03. This low false alarm rate is a highlight of the proposed method.

[Figure 6 near here]

We also computed empirical FDR across each time step for the 100 simulated data sets associated with each of the 19 change points starting at time $\tau = 5, 10, 15, \ldots, 95$. The empirical FDR was computed for each time step, and then for a specific changepoint, the mean empirical FDR was computed. Figure 7 displays the mean empirical FDR plotted against the change points. The figure shows that FDR is not sensitive to different change points.

[Figure 7 near here]

5. Application to 2011 German *E.coli* outbreak

The proposed algorithm was applied to the 2011 German *E.coli* O104:H4 outbreak. The data were obtained from the Robert Koch Institute in Germany (16). The outbreak recorded the highest number of Hemolytic Uremic Syndrome (HUS) cases reported in an outbreak. HUS can be classified as a food/waterborne disease. HUS causes excessive destruction of red blood cells;

the damaged blood cells clog the kidneys leading to fatal kidney failure. The *E.coli* O104:H4 outbreak affected 3,950 German residents; 800 of these suffered from HUS and 51 people died from their illness. The foodborne outbreak largely affected northern German states and continued from May to June of 2011.

All three statistics – Shewhart, CUSUM, and EWMA – were used with the proposed algorithm. The in-control data for each year corresponded to the data from the most recent year without an outbreak. For each state, bootstrap samples of size $B = 10{,}000$ were collected from the in-control period and used to estimate the p-values for each region at each time step.

The methodology was applied in a manner similar to the previous simulation study, with certain important choices related to the null distribution. The observed disease counts during in-control periods were frequently zero, so the mean disease count of the in-control data was often quite low. E.g., the sample mean of a year's worth of in-control data could be close to 0.3, but the maximum disease count during that period could be 7. Consequently, in order to make our analysis meaningful for detecting a true outbreak, the null mean value chosen for each testing period (μ_0 for the EWMA and λ_0 for the CUSUM method) was taken to be the maximum of the associated in-control period (typically, the data from the previous year); the null mean value of the bootstrap distribution was also shifted to the maximum of the associated in-control period.

Figure 8 shows the results for the state of Hesse. Figure 8 (a) shows the weekly disease counts for a 5-year period starting from January of 2006 to the end of 201l. The large spike on week 285 (orange line) signals the start of the outbreak on the second week of May 2011. Figure 8 (b) shows the three statistics over the same time interval and 7 (c) illustrates the alarm signals for all three methods. All three statistics – the Shewhart (blue), the EWMA (red), and the CUSUM (green) – detect the outbreak right at the onset on week 285. This is consistent with the simulation results shown in Figure 5, which illustrates how the detection speeds of all three statistics converge for relatively large shifts in mean disease counts, such as this spike on week 285.

In addition to displaying detection speeds, Figure 8 (c) also exhibits the extremely low occurrence of false alarms for all three methods. In fact, for the state of Hesse for all three methods, there was only one false alarm on week 66 over the entire five-year period from 2006 to 2011; in other words, there was only a single false alarm over 318 weeks. This result is consistent with the simulation results on Figure 6 illustrating the very low occurrence of false alarm rates. Figure 8 (c) was created to highlight this feature of the algorithm; therefore, the alarm signals for the entire five-year period were illustrated.

It is important to note the stepwise increment in the EWMA statistic (red) on week 107 in Figure 8 (b). For each year, the in-control means are updated based on data from the previous year. Recall that the most current EWMA statistic is calculated as the higher of the two statistics – the in-control mean or the current statistic. So, the reference line for the EWMA statistic is the in-control mean for each year, calculated from the previous year's data. Figure 8 (b) shows how the in-control mean for the EWMA statistic shifts up in year 2008 (starting with week 107) for the state of Hesse. However, the reference line for the CUSUM statistic is continuously zero, since the current CUSUM statistic is calculated as the higher of the two – the current CUSUM value or 0.

[Figure 8 near here]

In summary, out of the 16 German federal states, the outbreak occurred in 13 states. There was no significant increase in disease counts in the remaining 3 states – Rhineland Palatinate, Saxony, and Thuringia. The relatively faster EWMA and CUSUM statistics signalled alarms at the same time right at the onset of the outbreak in 5 of these states and a week later in the remaining 8. Similar to EWMA and CUSUM the Shewhart chart also detects the outbreak right at the onset in the same five states as these 5 states experience relatively large increases in disease counts. However, as the Shewhart chart is relatively slow, it detects the outbreak 3 weeks later in one state and 12 weeks later in another state.. The Shewhart chart fails to detect the outbreak in the remaining 6 states. In terms of false alarms, all three methods performed extremely well, signalling extremely few false alarms over the entire five-year surveillance period from 2006 – 2011.

6. Discussion

A new purely temporal multivariate method is proposed using modified statistical process control charts. Simulation studies were conducted using three charting statistics: the Shewhart, EWMA, and CUSUM statistics. In general, we recommend using the EWMA statistic when applying the proposed method because it is faster in detecting an outbreak compared to the other two methods.

The proposed method has certain features that are more appropriate for disease surveillance than most traditional multivariate SPC methods. First, the proposed method uses p-values to quantify strength of evidence for an outbreak as opposed to using conventional control limits, which only indicate whether an alarm should be signalled: specifically, the conventional methods do not indicate the strength of incompatibility between the observed data and the non-outbreak model. Second, the proposed method uses FDR for error control instead of conventional FWER-based methods. In a disease surveillance setting, an FDR of 0.05 would correspond to 5 false alarms out of a total of 100 alarms. However, with FWER based-methods, if the type I error is set to 0.05, this would signify that we expect 5 false alarms for every 100 tests during a non-outbreak period, which is not that relevant in a disease surveillance setting. Third, multivariate SPC methods that have already been proposed such as the method proposed by Li and Tsung (11) use regular FDR-controlling procedures such as the Benjamini and Hochberg (1) procedure. The proposed method uses a more powerful FDR control procedure – the ST procedure (21) – in order to increase the speed of detection and allow for possible correlation among the disease counts, which is critical in a disease surveillance setting. Fourth, the method is computationally simple, using multiple time series data, employing popular SPC charts, utilizing bootstrap resampling methods, and exploiting the easy to automate ST procedure for error control. Fifth, the proposed method – as demonstrated by extensive simulation studies – has rapid detection ability: this feature is critical in a disease surveillance setting, as the sooner an outbreak is identified, the earlier it enables health authorities to take preventive measures. Sixth, excessive false alarms rates are a common problem with conventional SPC charts. As noted by Buckeridge *et al.* (3), "EWMA and other SPC methods tend to produce false alarm rates in ranges that are not useful for public health practices." As illustrated in Figure 6 of section 3, the false alarms rates are mostly below 3% for all three statistics with the proposed method. In fact, it is a highlight of the proposed method. Seventh, the proposed method does not assume the disease counts to follow a particular distribution, unlike conventional SPC-based methods. The proposed method builds empirical in-control distributions using bootstrap methods. As "public health surveillance data tend to violate assumptions of SPC methods", as noted by Buckeridge *et al.* (3), the proposed method is more suitable for disease surveillance than the conventional methods. Eighth, the proposed method regularly updates baseline data to accommodate for changing population sizes, unlike conventional SPC methods that use data from a fixed in-control period.

Naturally, there are some limitations to the proposed method. While the proposed method is simple, it is not designed to deal with complex (e.g., seasonal or periodic) shifts in mean disease counts. Thus, while the method is appropriate for diseases that do not have any seasonality such as food and waterborne diseases like E. coli, it would need to be modified for monitoring diseases such as flu and asthma that have strong seasonality. Also, the method assumes that disease counts during the in-control period are independent. This is a common assumption with standard SPC methods that use industrial data as opposed to disease count data. However, as disease counts may have temporal correlation, the method can be further improved by accounting for that correlation structure.

A difference between real-life disease data and the data from the simulation study is that the simulation data had a persistent shift in the mean. Real-life disease patterns tend to autocorrelate due to natural factors such as weather, public health intervention, vaccines, etc. However, a major computational advantage of using persistent mean shifts in the simulation study is that signals cannot be totally missed, so detection timeliness comparisons can be made objectively without artificial penalties for undetected signals.

Acknowledgements

J. French was partially supported by NSF awards 1463642 and 1915277.

Conflict of Interest

The authors have declared no conflict of interest.

References

1. Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B 57(1995), pp. 289-300.

2. C.M. Borror, C.W. Champ, and S.E. Rigdon, *Poisson EWMA control charts,* Journal of Quality Technology 30(1998), pp. 352-361.

3. D.L. Buckeridge, H.S. Burkom, M. Campbell, W.R. Hogan and A. Moore, *Algorithms for rapid outbreak detection: a research synthesis*, Journal of Biomedical Informatics 38(2005), pp. 99-113.

4. M. Coory, S. Duckett, and K. Sketcher-Baker, *Using control charts to monitor quality of hospital care with administrative data*, International Journal for Quality in Health Care 20(2008), pp. 31-39.

5. Dassanayake, Sesha, and Joshua P. French. "An improved cumulative sum-based procedure for prospective disease surveillance for count data in multiple regions." Statistics in Medicine 35.15 (2016): 2593-2608.

6. Y. Elbert and H.S. Burkom, *Development and evaluation of a data-adaptive alerting algorithm for univariate temporal biosurveillance data*, Statistics in Medicine 28(2009), pp. 3226-3248.

7. S. Fasting and S.E. Gisvold, *Statistical process control methods allow the analysis and improvement of anesthesia care*, Canadian Journal of Anesthesiology 50(2003), pp. 767-774

8. R.D. Fricker, *Introduction to Statistical Methods for Biosurveillance,* Cambridge University Press, New York, 2013.

9. M. Frisen and C. Sonnesson, *Optimal Surveillance*, in *Spatial and Syndromic Surveillance for Public Health*, A.B. Lawson and K. Kleinman, eds., Wiley, West Sussex, 2005, pp. 31-52.

10. Lee, Sang-Ho, Jang-Ho Park, and Chi-Hyuck Jun. "An exponentially weighted moving average chart controlling false discovery rate." Journal of Statistical Computation and Simulation 84.8 (2014): 1830-1840.

11. Y. Li and F. Tsung, *Multiple Attribute Control Charts with False Discovery Rate Control,* Quality and Reliability Engineering Internatoinal 28(2012), pp. 857-871.

12. Z. Li, P. Qui, S. Chatterjee, and Z. Wang, *Using p values to design statistical process control charts*, Statistical Papers 54(2013), pp. 523-539.

13. J.M. Lucas, *Counted data CUSUMS*, Technometrics 28(1985), pp. 129-144.

14. D.C. Montgomery, *Introduction to Statistical Quality Control,* Wiley, New York, 2005.

15. E.S. Page, *Continuous inspection schemes*, Biometrika 41(1954), pp. 100-115.

16. Robert Koch Institute *Survstat@RKI database,* Retrieved September 17, 2018; from https://survstat.rki.de/Content/Query/Create.aspx

17. S.W. Roberts, *Control Chart Tests Based on Geometric Moving Averages*. Technometrics 1(1959), pp. 239-250.

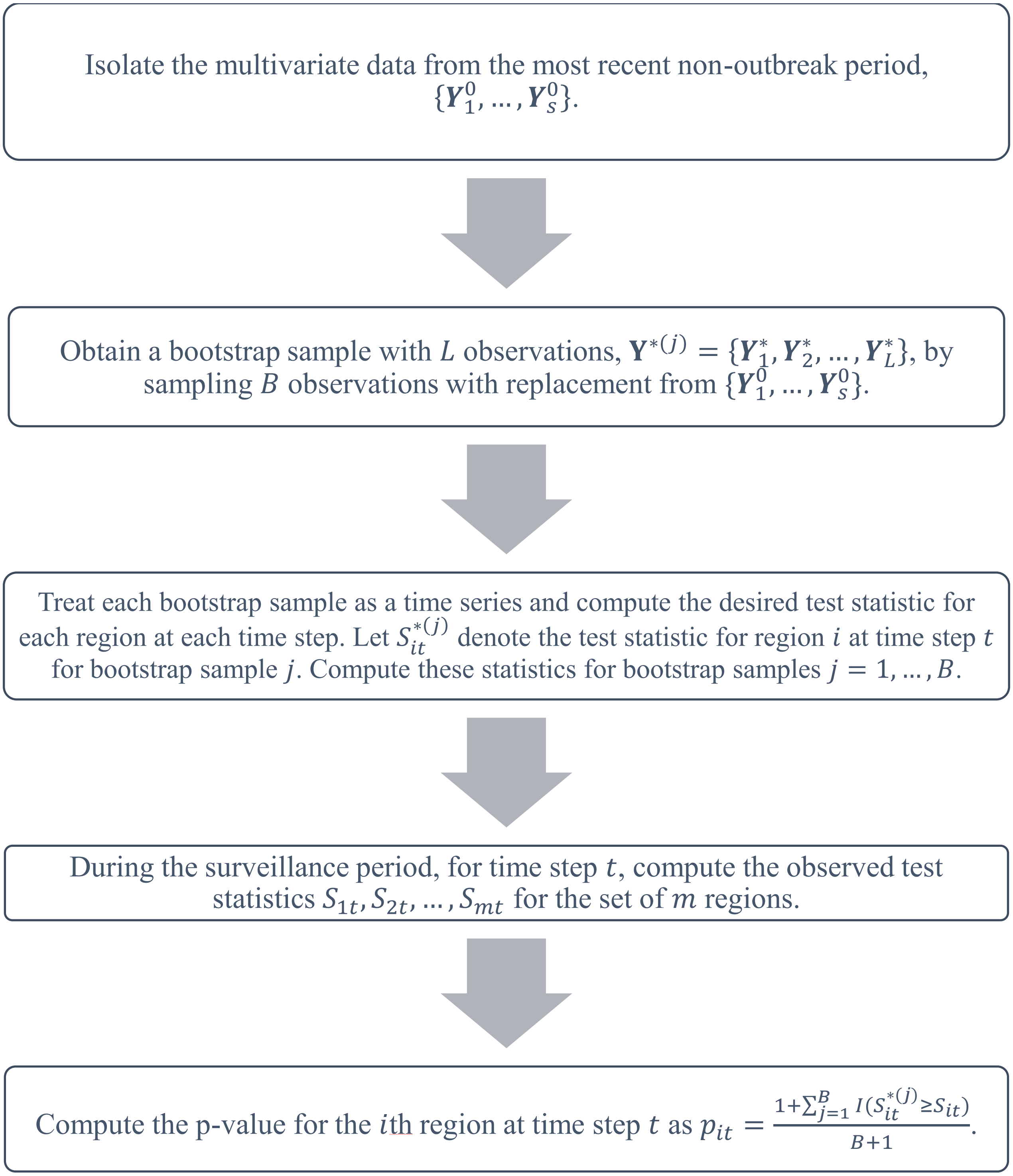
18. W.A. Shewhart, *Economic Control of Quality Manufactured Product*, D. Van Nostrand Company Inc., New York, 1931.

19. G. Shmueli and H. Burkom, *Statistical challenges facing early outbreak detection in biosurveillance*, Technometrics 52(2010), pp. 39-51.

20. L. Shu, S. Jiang, and S. Wu, *A one-sided EWMA control chart for monitoring process means*, Communications in Statistics - Simulation and Computation 36(2007), pp. 901-920.

21. J.D. Storey and R. Tibshirani, *Statistical significance for genomewide studies*. Proceedings of the National Academy of Sciences 100(2003), pp. 9440-9445.

22. W.H. Woodall, *Use of control charts in health care and public health surveillance (with discussion)*, Journal of Quality Technology 38(2006), pp. 88-103.



Isolate the multivariate data from the most recent non-outbreak period, $\{\boldsymbol{Y}_1^0, \dots, \boldsymbol{Y}_s^0\}$.

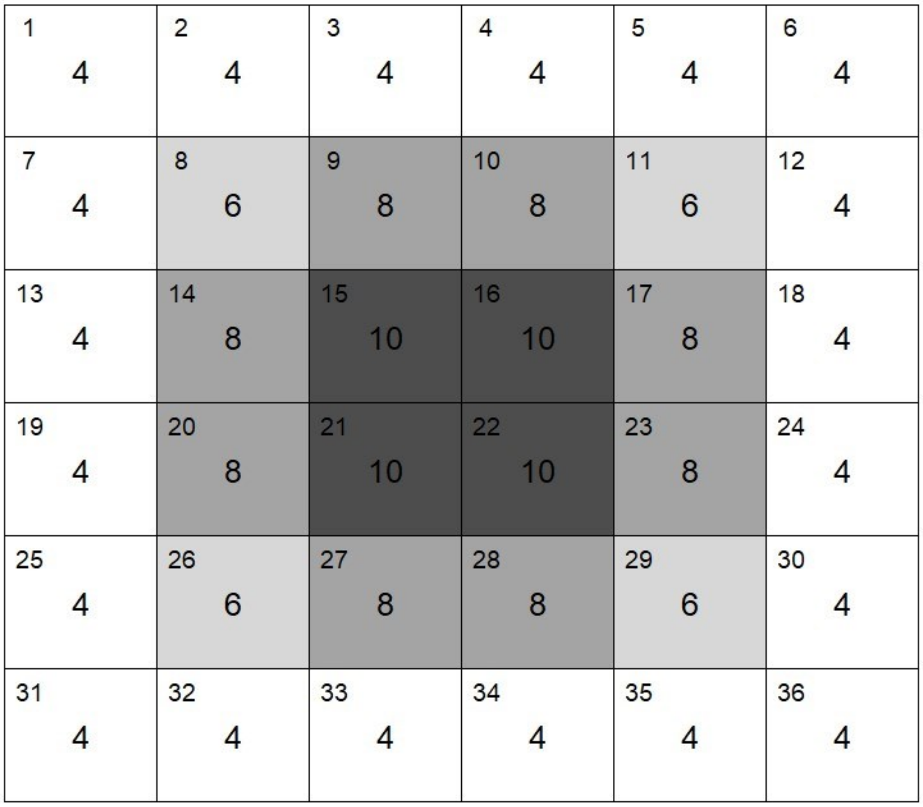
Obtain a bootstrap sample with L observations, $\mathbf{Y}^{*(j)} = \{\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*, \dots, \boldsymbol{Y}_L^*\}$, by sampling B observations with replacement from $\{\boldsymbol{Y}_1^0, \dots, \boldsymbol{Y}_s^0\}$.

Treat each bootstrap sample as a time series and compute the desired test statistic for each region at each time step. Let $S_{it}^{*(j)}$ denote the test statistic for region i at time step t for bootstrap sample j. Compute these statistics for bootstrap samples $j = 1, \dots, B$.

During the surveillance period, for time step t, compute the observed test statistics $S_{1t}, S_{2t}, \dots, S_{mt}$ for the set of m regions.

Compute the p-value for the ith region at time step t as $p_{it} = \frac{1+\sum_{j=1}^{B} I(S_{it}^{*(j)} \geq S_{it})}{B+1}$.

Figure 1. Flow chart describing the sampling and decision-making process.



1 4	2 4	3 4	4 4	5 4	6 4
7 4	8 6	9 8	10 8	11 6	12 4
13 4	14 8	15 10	16 10	17 8	18 4
19 4	20 8	21 10	22 10	23 8	24 4
25 4	26 6	27 8	28 8	29 6	30 4
31 4	32 4	33 4	34 4	35 4	36 4

Figure 2. The out-of-control mean disease counts ($\lambda_{1i},\ i = 1, \ldots, 36$) for each region is shown in the center and the region numbers are shown in the top left corner of each region.

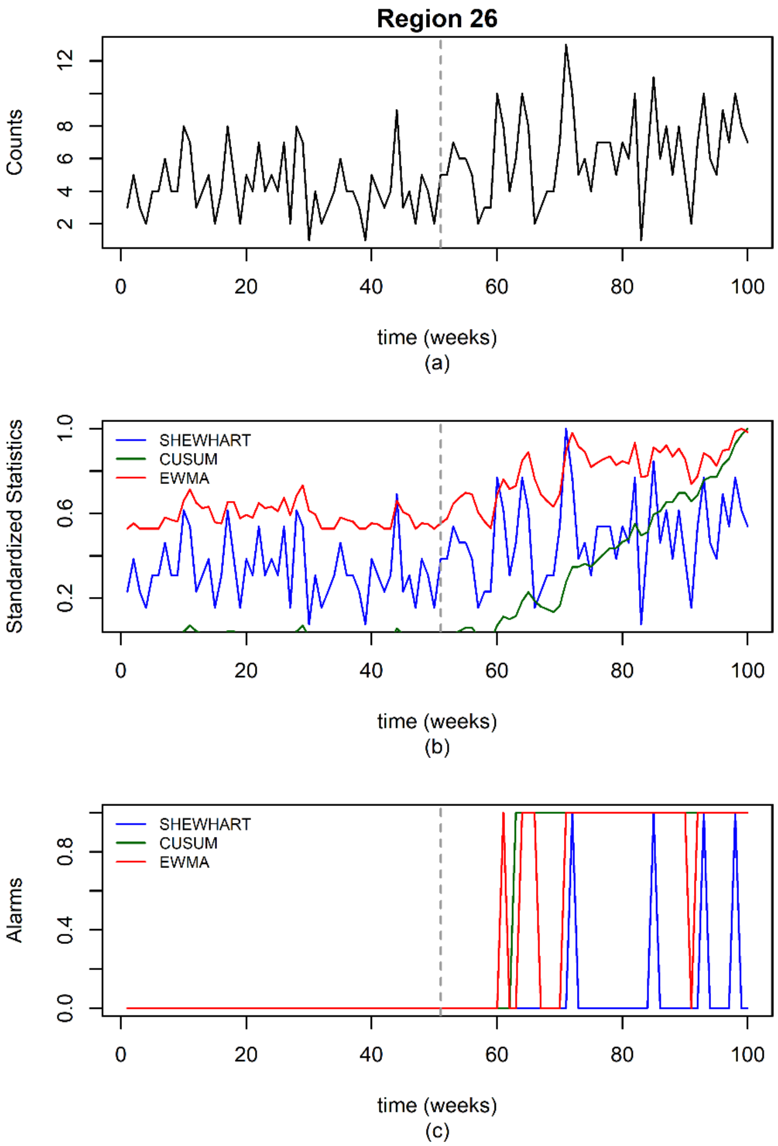


Figure 3. Summary plots for region 26 over time. Plot (a) shows the simulated weekly disease counts for days 1-100, plot (b) shows the standardized statistics – Shewhart (blue), CUSUM (dark green), and EWMA (red) – and plot (c) shows the alarms signalled by the Shewhart method (blue), CUSUM method (dark green), and EWMA method (red). The dotted grey line represents the start of the simulated outbreak period at week 51.

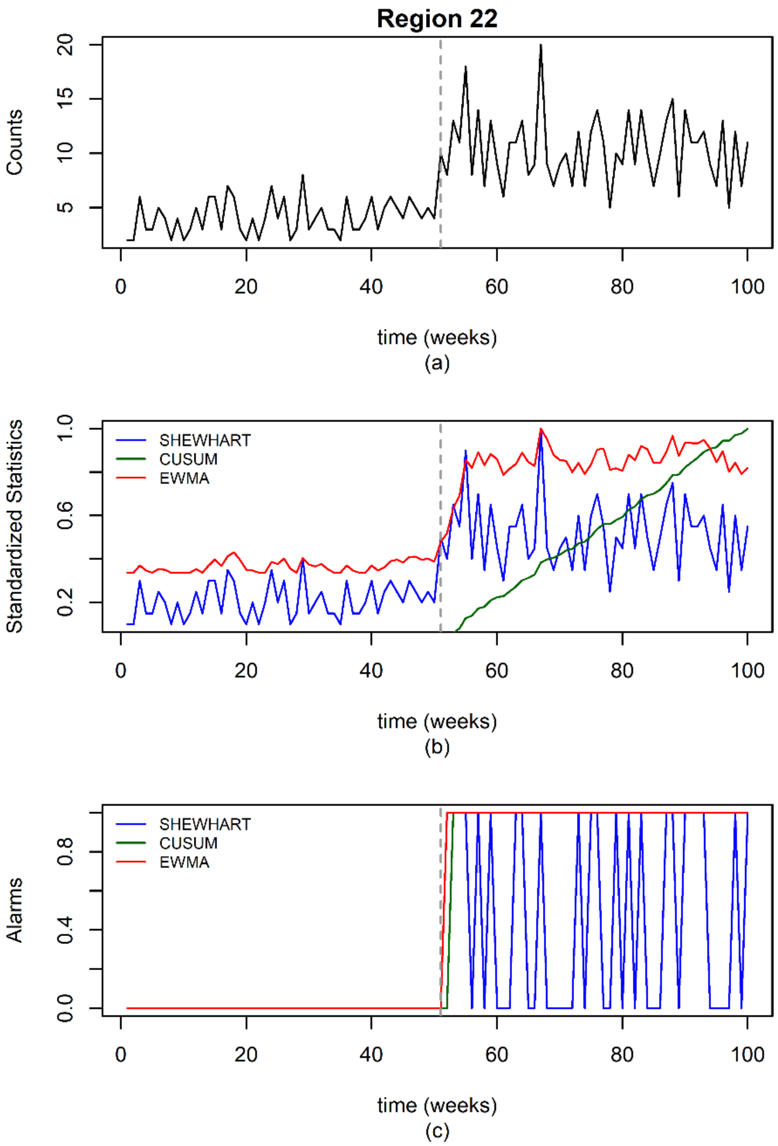


Figure 4. Summary plots for region 22 over time. Plot (a) shows the simulated weekly disease counts for days 1-100, plot (b) shows the standardized statistics – Shewhart (blue), CUSUM (dark green), and EWMA (red) – and plot (c) shows the alarms signalled by the Shewhart method (blue), CUSUM method (dark green), and EWMA method (red). The dotted grey line represents the start of the simulated outbreak period at week 51.

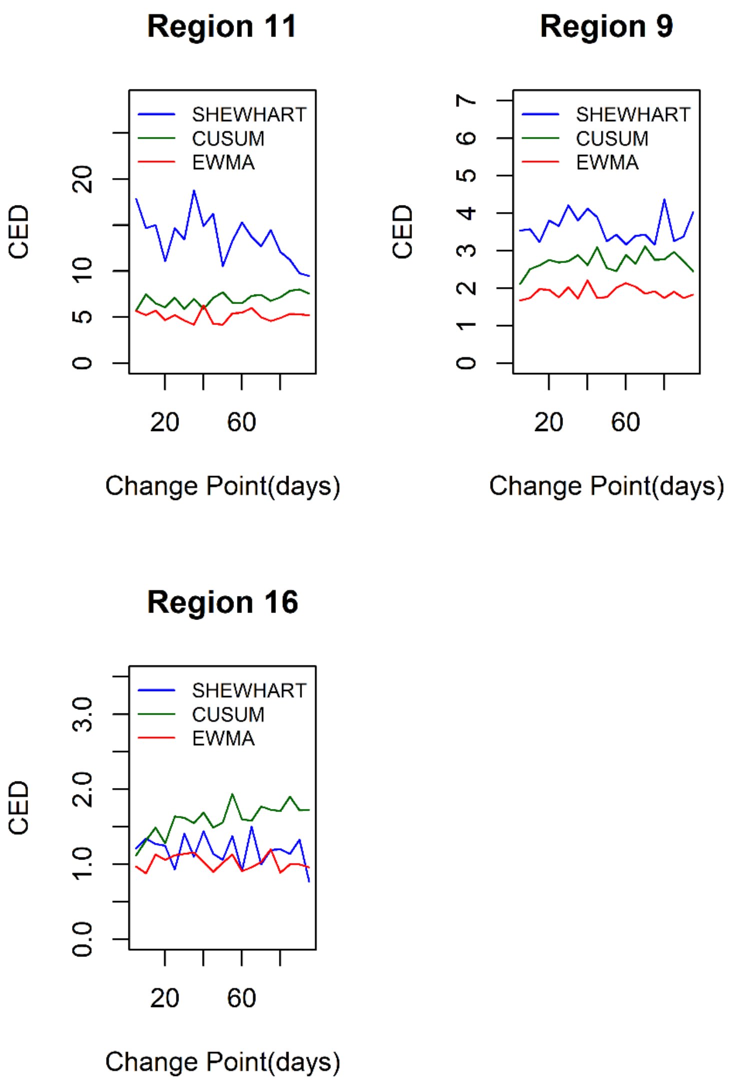
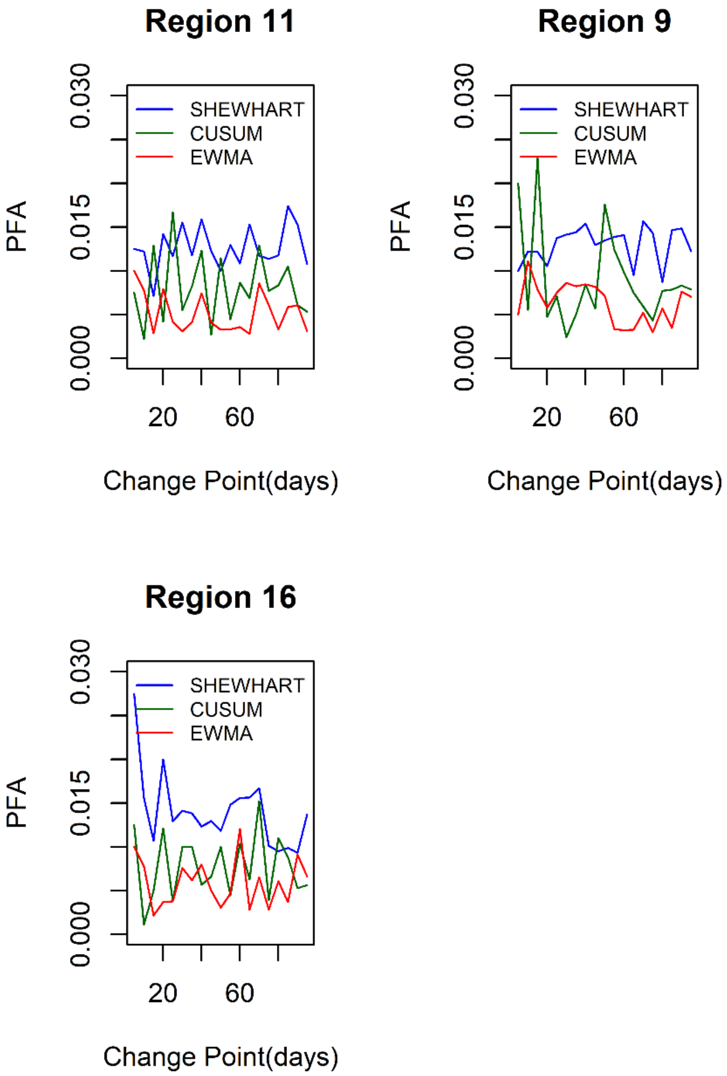


Figure 5. CED values versus change points for the three methods for regions 11, 9, and 16.



Region 11

Region 9

Region 16

Figure 6. PFA values versus change points for the three methods for regions 11, 9, and 16.

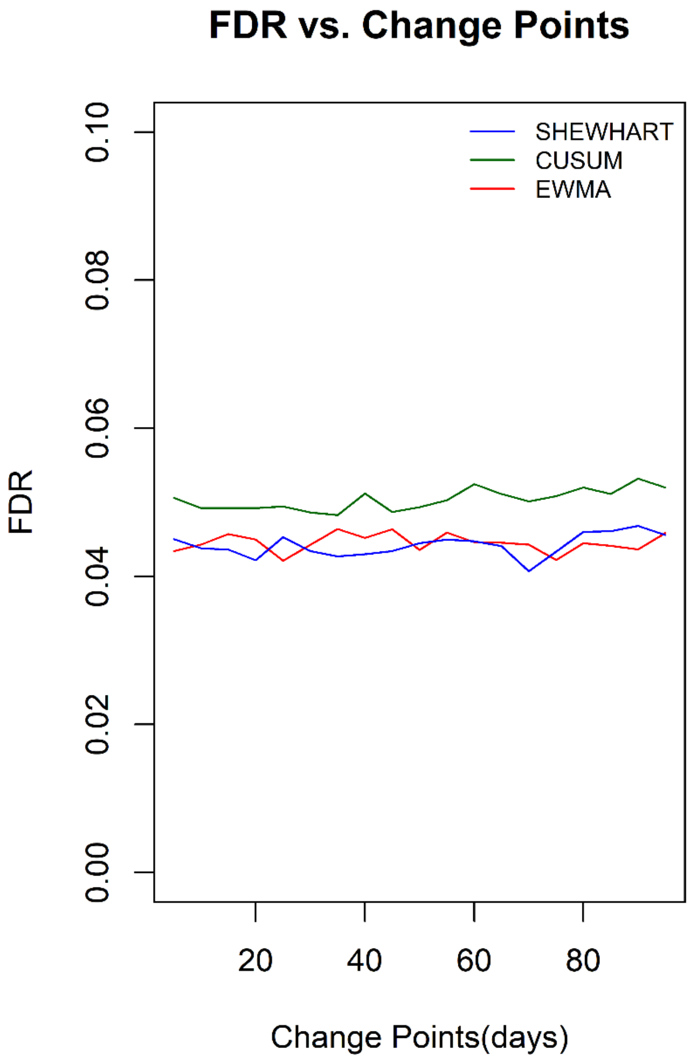


Figure 7. Mean empirical FDR of each step for the three methods for 100 data sets simulated with changepoints at $\tau = 5,10, \dots ,95$ days.

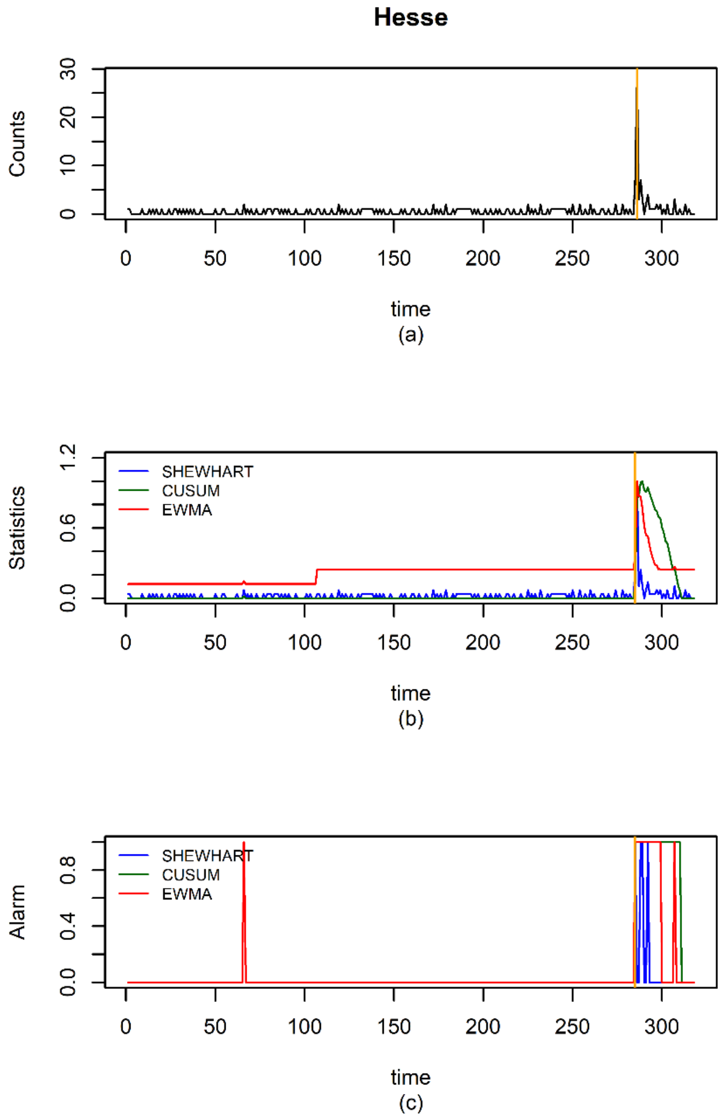


Figure 8. Weekly disease counts from 2006-2011 in the state of Lower Saxony are shown in (a). Shewhart (blue), CUSUM (dark green), and EWMA (red) statistics for the same period are shown in (b). Alarms signalled from the three methods – Shewhart (blue), CUSUM (dark green), and EWMA (red) - are shown in (c). The Orange line on week 285 signifies the start of the outbreak in this state.