Research Article

Statistics in Medicine

Received 16 March 2015, Accepted 05 January 2016 Published online 17 February 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6887

An improved cumulative sum-based procedure for prospective disease surveillance for count data in multiple regions

Sesha Dassanayake*† and Joshua P. French

We present an improved procedure for detecting outbreaks in multiple spatial regions using count data. We combine well-known methods for disease surveillance with recent developments from other areas to provide a more powerful procedure that is still relatively simple and fast to implement. Disease counts from neighboring regions are aggregated to compute a Poisson cumulative sum statistic for each region of interest. Instead of controlling the average run length criterion in the monitoring process, we instead utilize the FDR, which is more appropriate in a public health context. Additionally, *p*-values are used to make decisions instead of traditional critical values. The use of the FDR and *p*-values in testing allows us to utilize recently developed multiple testing methodologies, greatly increasing the power of this procedure. This is verified using a simulation experiment. The simplicity and rapid detection ability of this procedure make it useful in disease surveillance settings. The procedure is successfully applied in detecting the 2011 *Salmonella* Newport outbreak in 16 German federal states.

Keywords: disease surveillance; prospective disease surveillance; spatial CUSUM charts; multiple control charts; FDR

1. Introduction

Emerging disease clusters must be detected in a timely manner so that necessary remedial action can be taken to prevent the spread of an outbreak. Consequently, prospective disease surveillance has gained prominence as a rapidly growing area of research during the past decade [1]. In prospective surveillance, a decision is made at regular time intervals about the incidence of an outbreak, based on the data available up to that time. This is different from the traditional retrospective analysis where the entire data set is available for analysis.

Unkel *et al.* [1] broadly classify statistical methods proposed for prospective surveillance into time series methods, regression-based methods, statistical process control (SPC) methods, methods incorporating spatial information, and multivariate detection methods. Out of these methods, SPC methods have a long history of application in public health surveillance [2]. As the name suggests, many of the SPC methods originated in industrial process control but have lately been adapted for use in public health surveillance [3]. A key example is the cumulative sum (CUSUM) method, which is utilized in Bio-Sense, developed by the Center for Disease Control, and Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENSE), developed by the Department of Defense and Johns Hopkins University [4].

The CUSUM method cumulates the previous CUSUM statistic and the difference between an observed value and a reference value. The new statistic is the maximum of 0 and the cumulated value. When the process is 'in control' (i.e., no outbreak), we do not want to sound an alarm. When the process is 'out of control' (i.e., there is an outbreak), we want to sound an alarm. The method was originally developed for normally distributed data by Page [5] and was later extended to Poisson data by Lucas [6].

Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO, U.S.A.
**Correspondence to: Sesha Dassanayake, Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO, U.S.A.*
†E-mail: sesha.dassanayaka@ucdenver.edu

Raubertas [7] extended the purely temporal Poisson CUSUM method to a spatiotemporal setting. Disease counts from neighboring regions are pooled together to form regional neighborhoods. In pooling counts, a weighted average is calculated using the distance between regions as a weight. A CUSUM statistic is formed for each regional neighborhood, and an alarm is signaled when the CUSUM statistic exceeds the appropriate threshold. Later, Rogerson and Yamada [8] extended Raubertas's [7] method in two ways: First, they extended the method so that the expected counts can vary over time; second, they adopted a method that permits monitoring of regional neighborhood counts for *subregions* and their surrounding neighbors. The multiple testing problem that arises when testing multiple regional statistics simultaneously was handled by controlling the familywise error rate (FWER), using the popular Bonferroni correction. This method is somewhat similar to the method used in traditional multiple Poisson CUSUM charts utilized in industrial process control, where the multiple testing problem is handled by using the Bonferroni correction [9].

The objective of FWER procedures is to control the probability of even one false discovery, which exerts a stringent control when a large number of tests are performed simultaneously. Benjamini and Hochberg [10] popularized a less stringent error criterion for false discoveries, namely the expected proportion of false discoveries or the FDR. Compared with FWER-controlling methods, FDR-controlling methods have more power at the expense of additional false alarms. Benjamini and Hochberg [10] proposed a procedure for controlling the FDR in the context of independent test statistics. The procedure was extended to dependent statistics by Benjamini and Yekutieli [11]. FDR-controlling procedures have recently been adopted to handle multiple testing problems in SPC-based out-of-control processes [12].

Most CUSUM-based methods signal an alarm once the CUSUM statistic exceeds the appropriate threshold or critical value. In contrast, Li *et al.* [13] proposed a procedure using p-values, instead of critical values, to determine whether an alarm should be signaled. When the in-control distribution is known, Li *et al.* [13] suggest using Monte Carlo simulations to simulate the in-control distribution and estimate p-values. When the in-control distribution is unknown, bootstrap methods can be utilized to determine the empirical in-control distribution and estimate the p-values.

In the context of industrial process control, Li and Tsung [9] proposed a new CUSUM-based method that controls the FDR when multiple processes are being considered. One calculates a CUSUM statistic at each time step for each of the processes. The corresponding p-value for each process at each time step is calculated using a random walk-based approximation. Li and Tsung [9] provide the option of using the procedure of Benjamini and Hochberg [10] if the multiple statistics are assumed to be independent and the Benjamini–Yekutieli (BY) extension [11] if the statistics are assumed to be dependent.

The proposed method is a combination of several of these approaches: As proposed by Raubertas [7], the neighborhood disease counts are pooled in calculating the Poisson CUSUM statistic for each regional neighborhood; following the guidelines of Li *et al.* [13], these CUSUM statistics and bootstrap methods are used to compute p-values from which alarms are signaled—instead of using critical values. Adapting the method of Li and Tsung [9], FDR procedures are used to handle the multiple testing problem as opposed to using FWER-based techniques. We further improve the power of the procedure by utilizing an FDR-controlling technique proposed by Storey and Tibshirani [14] in addition to the method proposed by Benjamini and Yekutieli [11].

One begins the method by first setting an overall FDR level that the procedure is supposed to control. Then, at each time step, the disease counts of immediate neighbors are pooled together to compute a CUSUM statistic for each neighborhood. Next, Monte Carlo simulations or bootstrap methods can be used to approximate the in-control (no outbreak) distribution of the neighborhood CUSUM statistics, from which the corresponding p-values can be calculated. As multiple dependent statistics are tested simultaneously, two popular FDR-based methods are utilized to handle the multiple testing problem.

Section 2 provides additional details of the proposed method along with specifics of the Poisson CUSUM method and details of the popular FDR-based multiple testing procedures used in the algorithm. Section 3 provides the results of a simulation study using the proposed methodology, and Section 4 illustrates the results by applying the method to detect a *Salmonella* Newport outbreak in Germany; these results are compared with the results from traditional multiple CUSUM charts. Finally, Section 5 summarizes the strengths of the proposed method and provides some potential future directions for research.

2. Methods

2.1. Standard cumulative sum

The CUSUM method was proposed by Page [5] to detect small persistent changes in the mean of a process. The initial formulation of the CUSUM method assumed that the responses were uncorrelated and normally distributed. Later, Lucas [6] extended the CUSUM method to uncorrelated count data generated from a Poisson distribution. We describe the CUSUM method in this setting. Let $Y_1, Y_2, \ldots, Y_n$ be the response values at times $t = 1, 2, \ldots, n$. The goal is to sound an alarm when the response mean shifts from the 'in-control process' mean of λ_0 to the 'out-of-control process' mean of λ_1. The monitoring statistic used for the Poisson CUSUM method is

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

where Y_t is the count observed at time t, C_t is the Poisson CUSUM statistic at time t, k is a value chosen to minimize the detection time after a mean shift, and C_0 is defined to be 0. The CUSUM statistic is the larger of 0 and the sum of the CUSUM statistic for the previous time step with the difference between the observed value Y_t and the reference value k. The value k in equation 1 is chosen to minimize the time to detect a mean change from λ_0 to λ_1 (where $\lambda_1 > \lambda_0$) and is determined by the formula

$$k = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}\ . \tag{2}$$

An alarm is signaled when $C_t > h$, where h is a threshold or a critical value chosen to control the error rate. The threshold h is a function of the value of the parameter k, and the desired in-control average run length, ARL_0, is calculated using either Monte Carlo simulations or statistical tables. The ARL_0 is defined as the desired average time between two false alarms when the process is in control. Robertson *et al.* [15] point out that ARL_0 'can be difficult to specify.… In practice, approximations are used to estimate the value for h for a chosen ARL_0 [16], though this remains a key issue in CUSUM methods'. Moustakides [17] proved that the CUSUM was the optimal procedure for detecting the mean shift, in the sense that 'among all procedures with the same ARL_0, the optimal procedure has the smallest time until it signals a change, once the process shifts to the out-of-control state'. The popularity of the Poisson CUSUM method over other SPC methods is perhaps due to this theoretical optimality property.

2.2. Extending Poisson cumulative sum to a spatial setting

The original CUSUM method and its immediate extensions are univariate and purely temporal methods. Furthermore, different spatial extensions of the CUSUM methods have been proposed for spatiotemporal surveillance by Raubertas [7], Rogerson [18], and Rogerson and Yamada [8].

Raubertas [7] pooled the data in neighborhoods using distance-based weights to form regional neighborhoods. Suppose that we have m spatial regions where disease counts are observed, and let Y_{lt} be the disease count in region l at time t. Because larger populations are generally expected to have greater counts of disease incidence, an adjustment must be made before computing the value of k, previously described in equation 2. Let λ_{0l} be the expected number of cases in region l, n_l be the at-risk population in region l, and γ_0 be a baseline disease rate common to all regions. Then, we have $\lambda_{0l} = n_l\gamma_0$. If we desire to detect a disease outbreak when the disease rate shifts to γ_1, then $\lambda_{1l} = n_l\gamma_1$. Using λ_{0l} and λ_{1l}, the required k_l for calculating the CUSUM statistic can then be computed using equation 2. Although λ_{0l}, λ_{1l}, and the corresponding k_l can vary with changing population sizes, n_l, we considered the case of a constant fixed population in our simulation experiments, meaning k_l is fixed for the observed time period. The pooled count of unit i at time t is

$$Y'_{it} = \sum_{l=1}^{m} w_{il} Y_{lt},$$

where w_{il} is a measure of closeness between regions i and l. Similar pooling defines $\lambda'_{0i}, \lambda'_{1i}$, and k'_i. For each regional neighborhood, a pooled CUSUM statistic is calculated, and an alarm is triggered if the statistic exceeds a threshold. Extensions of this method allow time-varying parameters for k and λ [7]. Alternative approaches are given by Purdy *et al.* [19]. Our proposed methodology maintains the general framework of Raubertas [7], with important changes described later.

We also note that another possible approach for extending the Poisson CUSUM method to a spatial setting is to calculate a univariate CUSUM statistic for each region and then sound the alarm when the statistic exceeds a threshold adjusted to account for the multiple testing problem (e.g., adjusting using the Bonferroni correction). This is a method that is used in traditional multivariate process control [9]. Rogerson and Yamada [8] outline a similar method that can be applied for disease surveillance over multiple geographic regions. According to this method, the threshold h for each regional CUSUM chart is calculated as follows. First, the familywise type I error rate, α', is decided upon (e.g., a common choice is $\alpha' = 0.05$). Rogerson and Yamada [8] model run lengths as having an exponential distribution and desire the following relationship to hold: $p(\text{run length} < m) = \alpha'$, where m is the number of spatial regions. Using the exponential distribution, $p(\text{run length} < m) = 1 - \exp(-m\theta)$, where θ is the rate parameter of the exponential distribution and is chosen so that $1 - \exp(-m\theta) = 0.05$. This implies an average run length of $1/\theta$, because the mean of the exponential distribution is the reciprocal of the rate parameter. Using this ARL_0 and the region specific k, the threshold h for each region can be calculated. This method is applied to the case study presented in Section 4, where the results from this FWER-based method are compared with the results of the proposed FDR-based method.

2.3. *Modifications for a more powerful procedure*

Instead of using the traditional critical value-based testing method in the CUSUM procedure, we instead utilize p-values as described by Li *et al.* [13]. According to the first setting considered by Li *et al.* [13], when the in-control process distribution is known, p-values may be estimated using Monte Carlo simulations. More specifically, their algorithm starts by collecting in-control observations $Y_1, Y_2, \ldots, Y_t$ at a given time point t. With these observations, the corresponding CUSUM statistics $C_1, C_2, \ldots, C_t$ are computed. Next, B data sets from the in-control distribution are simulated, and for each of these data sets, the corresponding CUSUM statistics are computed. Using these statistics at each time point t, the empirical distribution of C_t is determined. At each time step, using C_t and the corresponding empirical distribution, the p-values are estimated as the proportion of samples where the associated CUSUM statistic is at least as large as the observed test statistic. When the in-control distribution is unknown, p-values may be estimated using bootstrap methods [13].

In addition to using p-values in our proposed procedure, we also utilize the FDR for error control. Because we will be monitoring multiple spatial regions, we clearly have a multiple testing problem. Li and Tsung [9] utilized the FDR to manage this problem in the context of traditional SPC methods using critical values; we follow the same pattern using p-values, instead of critical values.

Benjamini and Hochberg [10] popularized the concept of FDR to handle the multiple testing problem. The FDR is defined as the expected proportion of false discoveries among all discoveries:

$$FDR = E[V/R],$$

where V is the number of true null hypotheses that are declared significant and R is the total number of hypotheses that are declared significant out of m tests.

Numerous procedures have been proposed to control the FDR of numerous tests:

- In the simplest case, it is assumed that the m test statistics are independent. In that case, Benjamini and Hochberg [10] showed that the following procedure controls the FDR at level α. Consider testing hypotheses $H_1, \ldots, H_m$, and let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered observed p-values, with $H_{(i)}$ denoting the corresponding hypothesis. Define $j = \max\{i : p_{(i)} \leq \alpha i/m\}$, and reject $H_{(1)}, \ldots, H_{(j)}$. If no such i exists, reject no hypotheses. We will refer to this procedure as the Benjamini–Hochberg (BH) procedure.
- Benjamini and Yekutieli [11] proposed a modification of the BH procedure when the test statistics of the m tests are dependent. Specifically, if the $\alpha i/m$ utilized in the BH procedure is replaced with $\alpha / \sum_{i=1}^{m} i^{-1}$, then the FDR will still be controlled at level α. We will refer to this as the BY procedure. Both the BH and BY procedures were used by Li and Tsung [9] to control the FDR.
- A more powerful and up-to-date alternative is the Storey–Tibshirani (ST) procedure. Storey and Tibshirani [14] pointed out that the original FDR method proposed by Benjamini and Hochberg [10] is conservative and results in a substantial loss of power. To gain more power,

Storey and Tibshirani [14] introduced the q-value as an FDR-based measure of significance instead of the p-value, which is an FWER-based measure of significance. Whereas the p-value of a test measures the minimum false positive rate that is incurred when calling the test significant, the q-value of a test measures the minimum FDR that is incurred when calling a test significant. We describe this method in greater detail below.

An alternative definition of the q-value for a particular set of tests is that it is the expected proportion of false positives incurred when calling a test significant. Using this definition, Storey and Tibshirani [14] derived an expression for the false positive rate when calling all tests significant whose p-value is less than or equal to some threshold t, where $0 \leq t \leq 1$. The expected proportion of false positives is approximated by dividing the expected number of false positives by the expected total number of significant tests, for large m. The denominator is easily estimated by using the total number of observed p-values that are less than or equal to t. The numerator, the expected number of false positives, is much trickier and is obtained by multiplying the probability of a true null hypothesis being rejected by the total number of true null hypotheses, m_0. In finding the probability of a true null hypothesis being rejected, we need to use the result that the p-values corresponding to the null hypothesis are uniformly distributed. As a result, the probability that a null p-value being less than or equal to t is simply t. So the expected number of false positives is $m_0 t$. However, because m_0 is unknown, the proportion of true null hypotheses, $\pi_0 = m_0/m$, has to be estimated.

The key step in the method is an approximation of π_0. In estimating π_0, Storey and Tibshirani [14] make use of the fact that the distribution of true null hypothesis follows a uniform distribution and the distribution of truly alternative hypothesis is close to zero. They estimated π_0 by considering the density of the uniformly distributed section of the density histogram of the p-values.

Putting all of these components together, using the estimated proportion of true null hypotheses, $\widehat{\pi_0}$, the formula for computing the q-value, q_i, for a given p-value, t, is

$$\frac{\widehat{\pi_0}\, mt}{count\{p_i < t; i = 1, , , m\}}.$$

Likewise, a q-value for each test can be calculated, and the resulting q-values can be ordered at each time step from the smallest to largest. Once a significance threshold of α is decided upon, a set of significance tests is identified where a proportion of α is expected to be false positive.

It is important to make a distinction between FDR and an alternative quantity called pFDR, under which the q-value is technically defined. FDR can be loosely defined by $FDR = E(V/R)$, while the pFDR is defined as $pFDR = E(V/R|R > 0)$, to avoid V/R being undefined when $R = 0$ [14]. Just as a p-value is defined as the minimum false positive rate when calling the test significant, using pFDR, a q-value can be defined as the minimum pFDR at which the test is called significant [20]. For large m, when performing surveillance over a large number of smaller geographic units, such as for statewide or nationwide surveillance, $pFDR \approx FDR = E[V]/E[R]$.

As pointed out by Storey and Tibshirani [14], their procedure can be used for models with dependent statistics when the 'weak dependence' criterion is satisfied. According to them, 'as a rule of thumb, the more local the dependence is, the more likely it is to meet the weak dependence criterion'. The local dependence in the model generated by regional neighborhoods satisfies the weak dependence criterion. (For a formal mathematical definition of weak dependence, see remark D in [14].)

2.4. Detailed description of the proposed methodology

First, the desired FDR of all charts, α, is decided. At each time t, disease counts for each of the m regions $Y_{1t}, Y_{2t}, \ldots, Y_{mt}$ are collected. Disease counts of immediate neighbors are pooled to form regional neighborhoods with counts $Y'_{1t}, Y'_{2t}, \ldots, Y'_{mt}$. Then, the corresponding CUSUM statistics for the regional neighborhoods $C'_{1t}, C'_{2t}, \ldots C'_{mt}$ are calculated. For each region, B data sets are simulated based on the assumed known in-control distribution. Using these, for each regional neighborhood, an empirical distribution is determined from which the corresponding p-values $p'_{1t}, p'_{2t}, \ldots, p'_{mt}$ are calculated. Lastly, an appropriate FDR-controlling procedure is used to make decisions about whether an alarm should be signaled.

Alternatively, bootstrap methods can be used instead to calculate the p-values. As before, after defining the overall FDR level, at each time point t, disease counts for each of the m regions $Y_{1t}, Y_{2t}, \ldots, Y_{mt}$ are collected. Then, the disease counts of the immediate neighbors are pooled to form counts for regional

neighborhoods $Y'_{1t}, Y'_{2t}, \ldots, Y'_{mt}$. Using these counts, the corresponding CUSUM statistics for the regional neighborhoods $C'_{1t}, C'_{2t}, \ldots C'_{mt}$ are calculated. However, because the in-control distribution is unknown in this case, bootstrap methods are used to determine the empirical in-control distribution. Fixing the time period for which the counts follow the in-control distribution, B bootstrap samples of regional counts from the in-control time period are drawn with replacement across time while preserving the spatial relationships between the counts. Specifically, instead of sampling the counts of individual regions, we instead sample B times from the in-control time period. The entire set of regional counts for each sampled time is then used as a single sample of bootstrapped data. For the bootstrap samples, the corresponding CUSUM statistics are computed to determine the empirical in-control distribution. Using this empirical in-control distribution and the CUSUM statistics for the regional neighborhoods $C'_{1t}, C'_{2t}, \ldots C'_{mt}$, the corresponding p-values for each regional neighborhood at time t, $p'_{1t}, p'_{2t}, \ldots, p'_{mt}$, are calculated. Finally, a FDR-controlling procedure is used to determine the alarms, as before, because multiple p-values are compared simultaneously.

We will compare several FDR-controlling procedures in Section 3. The BH procedure will be used as a baseline. Next, the BY procedure will be used to see how accounting for dependence between test statistics improves performance. Lastly, the ST procedure will be used to highlight the additional gain in power.

2.5. *Contrast with existing methods*

The proposed method is better suited for disease surveillance compared with the conventional methods for several reasons: (i) the use of p-values in hypothesis testing is typically preferred over critical values; (ii) the use of FDR for error control is better in a public health setting than the standard ARL_0 used in SPC; and (iii) the multiple testing problem can be easily handled using a variety of off-the-shelf procedures. These points are elaborated below.

Conventional SPC methods were designed using thresholds or critical values. The use of p-values has many advantages over the conventional approach:

(1) P-values are used by most disciplines using statistics, whereas critical values are only used in certain contexts.
(2) When a signal of distributional shift is delivered, p-values can be used to quantify the strength of evidence of an outbreak. Results can indicate a range from no evidence, weak evidence, moderate evidence, strong evidence, and so on. However, with the conventional critical value approach, it is difficult to quantify the strength of evidence because the statistics are not on a standard scale from one setting to another.

The FDR is more appropriate for handling the multiple testing problem in surveillance scenarios than the ARL_0. The concept of ARL_0 originated in an industrial process control setting. When the process is in operation, an alarm is signaled when the CUSUM statistic crosses a predefined threshold, and the process is stopped. Sometimes, even when the process is in control, the CUSUM will signal an alarm. This is a false alarm and is analogous to a type I error. The ARL_0 is defined as the expected time until a false alarm. We make the following points in favor of using the FDR over ARL_0 in a disease surveillance setting:

The appropriateness of using FDR in a disease surveillance setting over the use of traditional ARL_0-based methods is elaborated upon later in the text:

(1) The FDR in a process-monitoring scheme may be regarded as the expected proportion of out-of control signals that turn out to be false alarms. On the other hand, the false alarm rate is the probability of concluding that the in-control process is out of control when it is actually in control. Therefore, FDR has a more meaningful interpretation in the surveillance setting than the false alarm rate. For example, if the FDR control level is 0.01, it is expected that one out of 100 alarms may be false. However, a false alarm rate of 0.01 simply means that we would sound an alarm for an in-control process one out of 100 times. We are more interested in controlling the error rate for out-of-control processes than in-control processes.
(2) FDR is directly related to predictive value positive (PVP), one of the seven attributes used by the Center for Disease Control for evaluating surveillance systems, because $FDR = 1 - PVP$.
(3) The ARL_0 does not make sense in a health context although it is meaningful in an SPC context. A run length is defined as the number of observations from the starting point to the point where the

statistic crosses a predefined threshold. When such a shift occurs, the process might be stopped and restarted again. In contrast to an industrial process, an emerging disease outbreak cannot be stopped, so the concept of a run length is not as appropriate in a health context.

Lastly, using FDR provides us with a host of up-to-date methods to handle the multiple testing problem such as BH [10], BY [11], and ST [14] procedures. It is possible to gain more power with these methods compared with techniques controlling the FWER such as the Bonferroni correction.

3. Simulation experiment

3.1. Experiment setup

Our simulation experiment consists of a study area composed of 25 regions on a five-by-five grid. The time period considered by the simulation experiment was 100 time points or 'days'. No outbreak occurred during the first 50 days (days 1–50); disease counts for each of the 25 regions for the first half of the simulation were simulated as independent Poisson random variables with a constant mean of $\lambda_{0i}=4,\ i=1,\ldots,25$. However, for the second half of the simulation (days 51–100), an outbreak was simulated where the Poisson means for the regions peak in the center and tapered toward the edges. The out-of-control means are denoted by $\lambda_{1i},\ i=1,\ldots,25$.

Specifically, as illustrated in Figure 1, the four corner regions 1, 5, 21, and 25 obtain the mean shifts of 0.2 standard deviations from the original mean. Regions 2, 3, 4, 6, 10, 11, 15, 16, 20, 22, 23, and 24 obtain 0.3 standard deviation mean increases to 4.6. The middle regions 7, 8, 9, 12, 14, 17, 18, and 19 obtain 0.75 standard deviation mean increases to 5.5. Finally, the mean of region 13 increases by one standard deviation to 6. The out-of-control process means are shown in Figure 1.

For this simulation, each of the 25 CUSUM statistics was designed with an in-control Poisson mean (λ_0) of 4 and an out-of-control Poisson mean (λ_1) of 6. In other words, because the in-control mean is 4 (standard deviation is 2), the CUSUM statistic is built to detect a change of one standard deviation increase. Using the first method of Li *et al.* [13], 10,000 Monte Carlo simulations were used to determine the empirical in-control distribution. The desired level of the FDR was set to $\alpha=0.05$.

3.2. Simulation results

Figures 2 and 3 show the results for a single simulation for regions 1 and 13, respectively. Region 1 has the smallest shift in mean with a 0.2 standard deviation increase while region 13 has a one standard deviation increase.

Figure 2 summarizes the results of a single simulation over the 100-day period for region 1, where the process mean was 4 for times 1–50 and 4.4 for times 51–100. Figure 3(a) shows the disease counts for

1 4.4	2 4.6	3 4.6	4 4.6	5 4.4
6 4.6	7 5.5	8 5.5	9 5.5	10 4.6
11 4.6	12 5.5	13 6	14 5.5	15 4.6
16 4.6	17 5.5	18 5.5	19 5.5	20 4.6
21 4.4	22 4.6	23 4.6	24 4.6	25 4.4

Figure 1. The out-of-control mean disease counts (λ_{1i}, $i=1,2,\ldots,25$) of the 25 regions for days 51–100 of the simulation experiment are indicated in the center of the regions. Region numbers are indicated at the top of each region in smaller font.

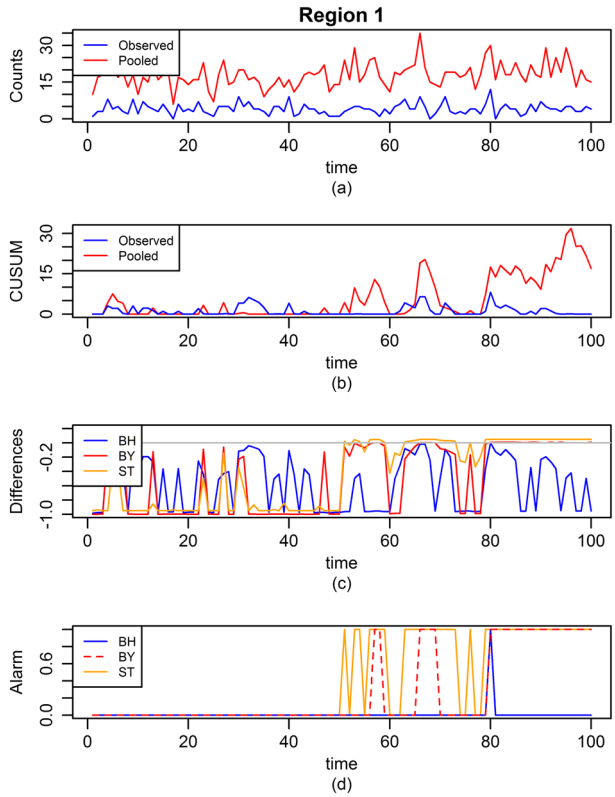

Figure 2. Summary plots for region 1: (a) disease counts, (b) CUSUM statistics, (c) differences, and (d) alarms. In plot (c), the blue line shows the difference, ($\alpha_{BH} - p$-value), where $\boldsymbol{\alpha_{BH}}$ is the threshold using the BH procedure for the independent model. Similarly, the red line shows the difference, ($\alpha_{BY} - p$-value), where $\boldsymbol{\alpha_{BY}}$ is the threshold using the BY procedure for the pooled model; the orange line shows the difference, ($0.05 - q$-value), where 0.05 is the overall FDR using the ST procedure for the pooled model. The gray horizontal line indicates a zero difference.

region 1 over the 100-day period for a single simulation. The blue line shows the observed counts in region 1, whereas the red line shows the pooled counts: Pooled counts were calculated by totaling the counts of the region of interest, in this case region 1, and its immediate neighbors that share a common boundary (regions 2, 6, and 7). In other words, the region of interest and its immediate neighbors each obtain a weight of 1. Figure 3(b) shows the CUSUM statistic for the observed (blue) and pooled (red) counts; note that the CUSUM statistic for the observed counts does not show any large increase, except for a relatively small spike at day 80, in the presence of the outbreak (days 51–100). Recall that region 1 obtains the smallest increase of a 0.2 standard deviation shift in the mean. The CUSUM statistic using the observed regional counts is designed to respond to a much larger increase of a one standard deviation shift and is not sensitive to this relatively small change. However, the CUSUM statistic for pooled regional neighborhood counts captures information from neighboring counts as well and indicates a substantial rise after day 51. In Figure 3(c) the blue line shows the difference, ($\alpha_{BH} - p$-value), for the independent model, where α_{BH} is the p-value threshold using the BH procedure; similarly, the red line shows the difference, ($\alpha_{BY} - p$-value), for the pooled model where α_{BY} is the p-value threshold using the BY procedure; finally, the orange line shows the difference, ($0.05 - q$-value), for the pooled model, where 0.05 is the overall FDR at each time using the ST procedure. All three plots are displayed together for ease of comparison. Note that the point at which an alarm will be signaled is the time at which the difference exceeds 0, shown by the gray horizontal line. The last figure, Figure 3(d), displays these alarms for the three multiple comparison procedures previously discussed: BH, BY, and ST. The BH procedure (applied to the CUSUM statistic from the observed counts) sounds only one alarm on day 80. The BY procedure (applied to the CUSUM statistic for pooled counts) starts to sound alarms after day 57, while the ST procedure starts indicating regular alarms much earlier (beginning at day 51) and always sounds an alarm when the BH procedure sounds an alarm.

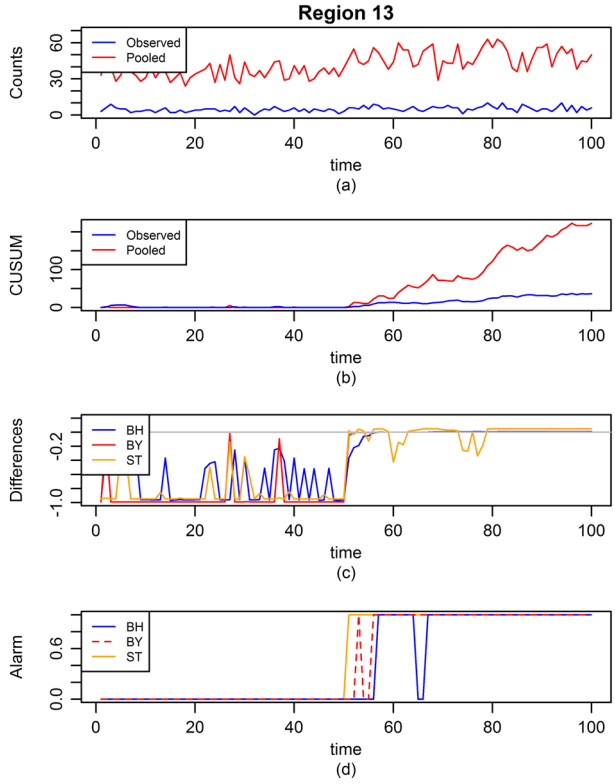

Figure 3. Summary plots for region 13: (a) disease counts, (b) CUSUM statistics, (c) differences, and (d) alarms. In plot (c), the blue line shows the difference, ($\alpha_{BH} - p$-value), where $\boldsymbol{\alpha_{BH}}$ is the threshold using the BH procedure for the independent model. Similarly, the red line shows the difference, ($\alpha_{BY} - p$-value), where $\boldsymbol{\alpha_{BY}}$ is the threshold using the BY procedure for the pooled model; the orange line shows the difference, ($0.05 - q$-value), where 0.05 is the overall FDR using the ST procedure for the pooled model. The gray horizontal line indicates a zero difference.

Figure 3 contains the plots for region 13, with the highest step increase of one standard deviation. Note that the CUSUM statistic in this simulation was designed to detect a change of this magnitude. The alarms are more persistent for all three procedures, with the ST procedure (orange line) identifying the outbreak the earliest (on day 51, right at the beginning of the outbreak), followed by the BY procedure (red line) on day 53 and the BH procedure (blue line) on day 57.

To confirm that the three procedures are controlling the FDR at the desired level of $\alpha = 0.05$, we calculated the observed proportion of false discoveries for each procedure for 100 independent simulations, using the same experimental setup in Section 3.1 for each simulation. Box plots of the results for each procedure are displayed in Figure 4. Clearly, the overall FDR is below the specified 0.05 for all three procedures.

3.3. Additional investigation of power

We now discuss the 'power' of the three procedures more carefully. To show that these results are consistent, the conditional expected delays (CEDs) and the probability of alarm (PAs) for a large number of simulations were calculated for the three models.

The CED and probability of a false alarm (PFA) are commonly used measures of evaluation in surveillance. Frisén [21] discusses these two measures and several performance measures used in statistical surveillance. CED is a measure commonly used to compare speed of detection, and PFA is used to compare false alarm rates.

Let $t_A = \min\{t : S(t) > t\}$ be the alarm time of a surveillance statistic crossing the threshold h. CED is the average delay time until an alarm when the change occurs at time point τ and defined as

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

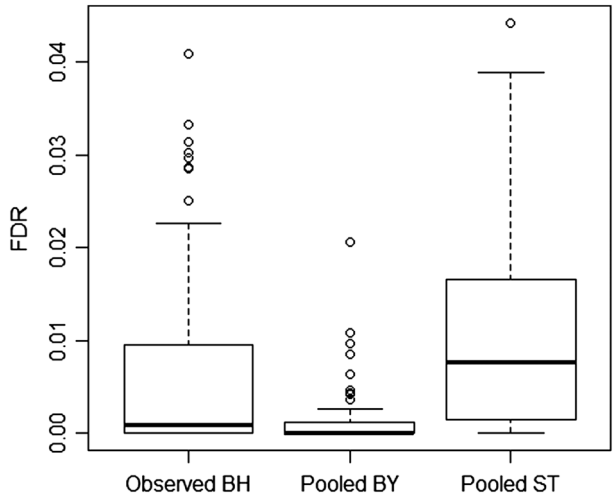

Figure 4. The box plots of proportion of false discoveries for all three procedures over 100 simulations. Results for the BH procedure are shown on the far left, those for the BY procedure in the middle, and those for the ST procedure on the far right.

The other measure, PFA is defined as

$$PFA = P(t_A < \tau).$$

Robertson *et al.* [15] point out that a surveillance system should provide 'quick detection and few false alarms'. Furthermore, Jiang *et al.* [22] point out that 'design and evaluation of a surveillance method need to trade-off between false alarm probabilities and detection delays'. Therefore, PFA and CED are useful measures to evaluate the performance of surveillance systems.

In our setting, we are controlling the FDR. Because the FDR should be held constant across procedures, a more appropriate measure of power in our context is simply the PA. Because the FDR at each time step is held constant across procedures, the most powerful procedure will be the one that sounds an alarm most frequently; we estimate this by dividing the total number of alarms by the total number of tests.

In calculating CED and PA for the three models, a range of change points (5, 10, 15, …, 95) was considered (recall that the simulation results presented earlier were for the case where the change point was day 51). For each change point, 100 simulations were performed. For example, 100 data sets were generated where the change point for the outbreak was at day 5. Similarly, 100 data sets each were generated for the other change points such as days 10 and 15. Then, for each change point, the CEDs of the 100 simulated data sets were averaged for each of the three models, respectively. In order to avoid issues in estimating the CED for change points near the upper end of the original simulation time period (1, 2, …, 100), the time period under consideration was extended up to 150 so that a change could be detected for all data sets.

Figure 5 shows the CED plots for the BH procedure (blue), the BY procedure (red), and the ST procedure (orange) for regions 1, 2, 7, and 13. Recall that the increase in the out-of-control process mean for each region is different, with region 1 having the smallest increase and region 13 having the largest.

The top left plot in Figure 5 shows the average CED against change point time for region 1 for all three models. Clearly, for all four regions, the ST procedure has the lowest CED, followed by the BY and BH procedures, respectively, across all change points. In general, as evident by the four plots in Figure 5, the procedures using the pooled regional neighborhood counts (models using BY and ST procedures) signal much faster than the model that only uses observed regional counts (model using the BH procedure); within the pooled models, the ST procedure detects outbreaks faster than the BY procedure.

We now compare the performance of the procedures using PA. Figure 6 shows the PA against change points for all three procedures in regions 1, 2, 7, and 13. Again, the same choice of change points (5, 10, 15, …, 95) were considered. For each change point, the PAs for 100 simulations were averaged for all three methods. As expected, in all four regions, the ST procedure has the highest PA followed by the BY

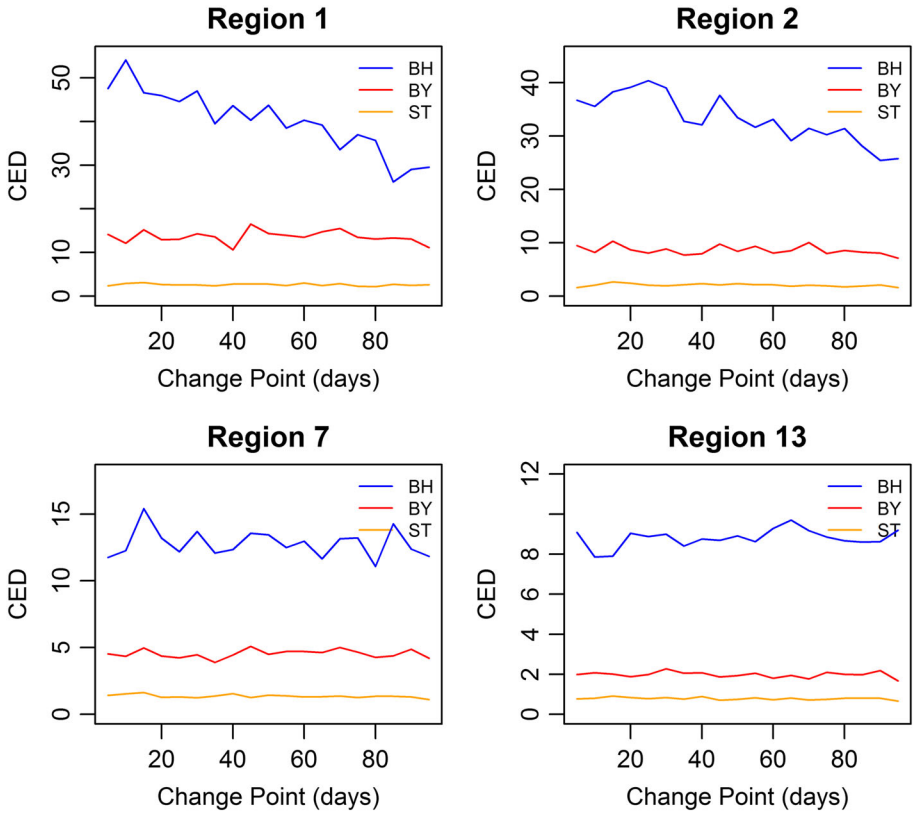

Figure 5. Average CED versus change point time for BH, BY, and ST models for regions 1, 2, 7, and 13.

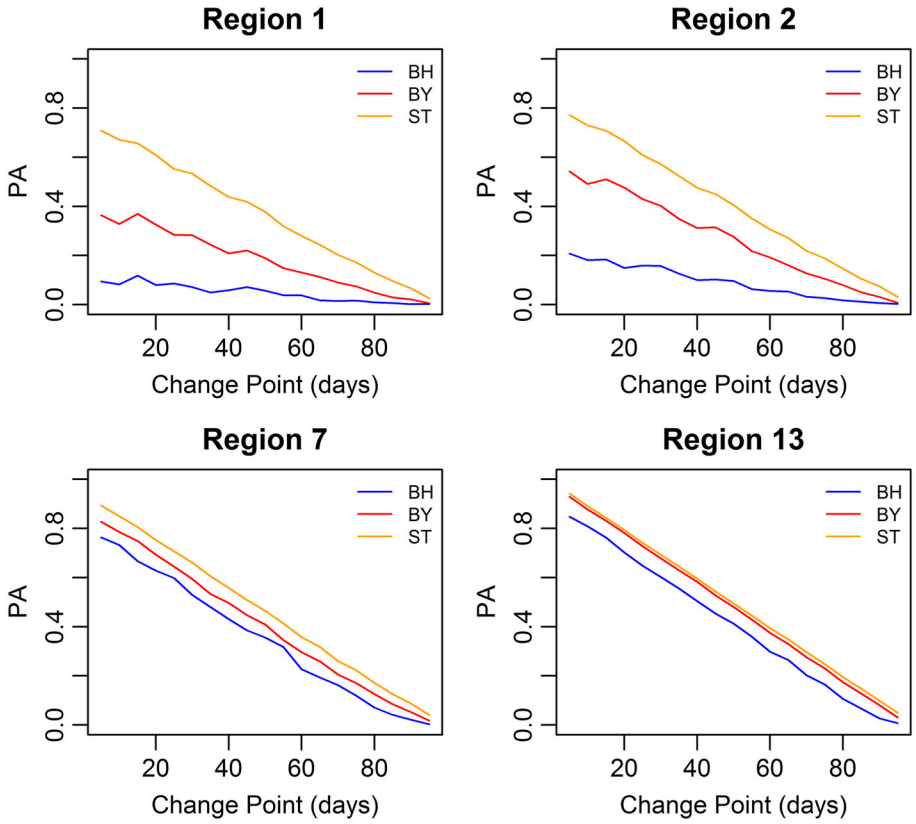

Figure 6. Average probability of an alarm versus change point time for BH, BY, and ST models for regions 1, 2, 7, and 13.

procedure and then the BH procedure. However, for larger increases in the mean, the vertical distance between the three lines tends to diminish, indicating that the gains in power tend to be smaller for larger step increases in the mean. Because FDR is controlled at a 0.05 level, the higher the total number of alarms, the higher the number of true alarms. So when FDR is controlled, we want to sound the alarm more frequently. As expected, the ST procedure (orange) sounds the most alarms followed by the BY procedure (red) and finally the BH procedure (blue). The downward trends in the plots are due to the following reason: When the outbreak occurs earlier during the 100-day period, the duration of the outbreak is longer, allowing more chances to detect the out-of-control process.

4. Application to *Salmonella* data

The proposed method was applied to a data set of *Salmonella* Newport cases reported weekly from 16 German federal states between year 2004 and year 2014. The data were obtained from the Robert Koch Institute in Germany [23]. Because *Salmonella* is not a contagious disease, without trend or seasonality, the counts were assumed to be independent. The 16 German federal states are displayed in Figure 7.

The first 2 years of data (2004–2005) were used to estimate the in-control distribution in each state because there were no unusually high disease counts reported from any of the states during this period. A Poisson dispersion test [24] was used to ensure that the Poisson counts were not overdispersed. No overdispersion was detected in the first 2 years of data at a type I error rate of 0.01. Also, it is reasonable to assume that the data are spatially independent over the first 2 years, which we use as the in-control time period.

The proposed surveillance model was applied to data from 2006 to 2013. Because states with larger populations are likely to have larger disease counts, the method proposed by Raubertas [7] detailed in Section 2.2 was adopted to address this issue. For the pooled model, disease counts from each state were pooled together with those of the immediate neighbors, using a binary connectivity matrix similar to the one used in the simulation experiment. For the pooled counts, the CUSUM statistic was computed, the *p*-values for each region were estimated, and an alarm was signaled using an FDR-controlling procedure.

Figure 7. Map of the 16 federal states of Germany.

P-values were estimated using the bootstrap method because the in-control distribution was unknown. The BH procedure was used to handle the multiple testing problem for the independent counts model while the ST procedure was used for the pooled model; for both models, the overall FDR was set to 0.05 to minimize false alarms.

A plot of the results for both the independent counts and pooled counts models are shown in Figures 8 and 9 for the states Bavaria and Saxony-Anhalt, respectively.

The data set contained counts for 528 weeks starting from the first week of January in 2004 to the second week of February in 2014. The time axis shows the number of weeks, starting from the first week of January 2004 (week 1). Because the *Salmonella* Newport outbreak actually occurred in 2011 [25], the results are plotted from the last week of September 2009 (week 300) to the second week of February 2014 (week 528). The results indicate that the pooled model using the ST procedure was successful in detecting the *Salmonella* outbreak in the first week of November 2011 (week 410), and this was consistent for all 15 states (The pooled model using the BY procedure also gave alarms on week 410 in all 15 states as well, as the rise in disease counts was fairly large.) In contrast, the model using independent state counts was rather inconsistent in signaling an alarm around the period of the outbreak: The model did not detect an outbreak in Bremen, which has the smallest population; in Bavaria, the outbreak was detected 53 weeks later; in Rhineland-Palatinate and Saxony-Anhalt, it was detected 2 weeks later; in Baden-Württemberg and Schleswig-Holstein, it was detected a week later.

These results using the proposed method employing FDR techniques can be compared with the performance of traditional multiple Poisson CUSUM methods using FWER-based multiple testing procedures. The traditional method using 15 univariate Poisson CUSUM charts with a Bonferroni correction produced rather inconsistent results in signaling the alarm: In Bremen and Saxony, no change was detected; in Bavaria, the outbreak was detected 53 weeks later; in Baden-Württemberg, it was detected 4 weeks later; in Rhineland-Palatinate, Schleswig-Holstein, and Thuringia, the outbreak was detected 2 weeks later; in Mecklenburg-Vorpommern, it was detected a week later.

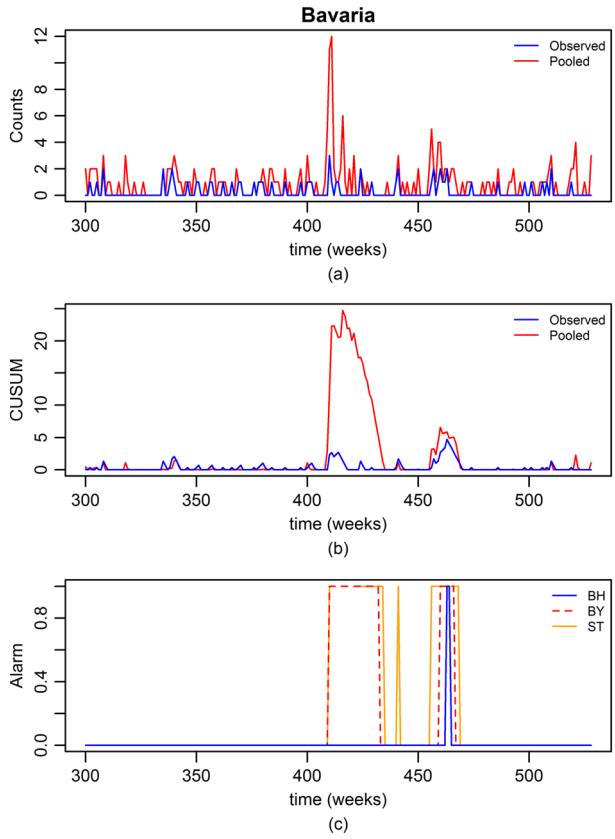

Figure 8. Pooled disease counts (a), CUSUM statistics (b), and alarm plots (c) for Bavaria. The blue line in (a), (b), and (c) represents the independent-count model, and the red line represents the pooled-count model for Bavaria.

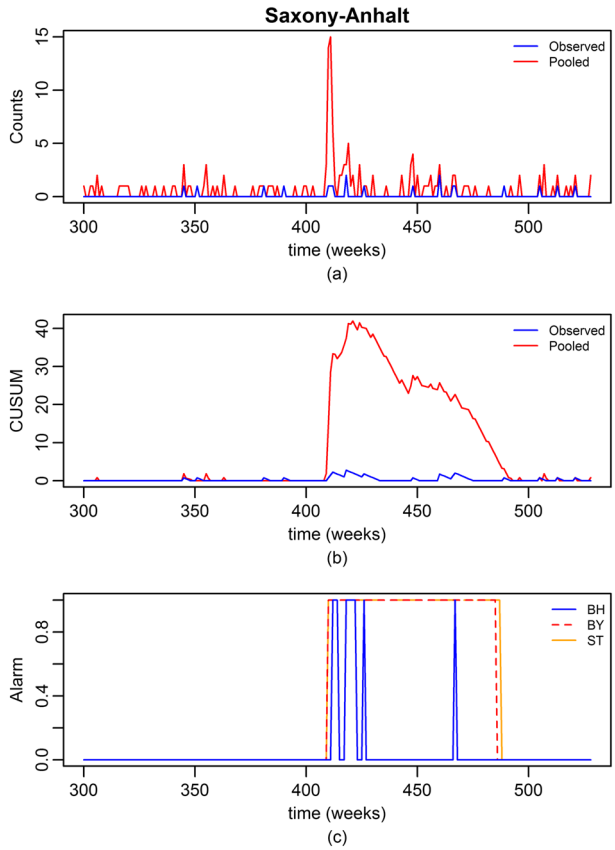

Figure 9. Pooled disease counts (a), CUSUM statistics (b), and alarm plots (c) for Saxony-Anhalt. The blue line in (a), (b), and (c) represents the independent-count model, and the red line represents the pooled-count model for Saxony-Anhalt.

The alarm plot on Figure 8(c) shows the results for Bavaria: Using the pooled model with the ST procedure, the outbreak is detected in week 410, consistent with the results of the other states, but there is a much longer delay (53 weeks) in detection using the independent counts model. Because the increase in disease counts is relatively large in week 410, both ST and BY procedures detect the outbreak in week 410. However, note that there is another relatively smaller increase in counts after week 450. This smaller shift is detected first by the more powerful ST procedure in week 452, followed by the BY procedure in week 456 and then the BH procedure in week 459. Although the ST and BY procedures are comparable in terms of detection speed when it comes to detecting large increases, the ST procedure is quicker in detecting relatively smaller changes, consistent with the simulation results.

Figure 9(c) shows the results for Saxony-Anhalt: Using the pooled model with the ST procedure, the outbreak is again detected in week 410, but there is a 2-week delay in detection using the independent model. The pooled model sounds a steady alarm after week 410 for a longer period; however, the independent model sounds an alarm irregularly for a relatively shorter period. In summary, the pooled model using the ST procedure was successful in detecting the outbreak simultaneously throughout the country as opposed to the independent model, which signaled alarms inconsistently and following considerable delays.

5. Discussion

In the proposed new procedure, regional disease counts from multiple regions are aggregated to compute an alarm statistic using the popular Poisson CUSUM method. Two novel aspects of the proposed method are particularly advantageous in a disease surveillance setting. First, the use of p-values (instead of the commonly used critical values) enables us to evaluate the strength of the out-of-control signal. Second, the use of FDR for error control instead of the standard FWER or ARL_0 allows the use of powerful tools to handle the multiple comparison problem. The ST [14]

procedure was the most powerful procedure in our tests, in comparison with the more conservative BH [10] and BY [11] procedures. The simplicity of the algorithm and the improved speed of detection make the proposed method useful in a practical disease surveillance setting, as illustrated by the *Salmonella* Newport example in Germany.

We emphasize that the proposed procedure controls the FDR at each time step. If one wanted to control FDR over all time steps simultaneously, the BY and ST procedures would still apply, although the procedure would no longer be prospective. As evident from simulation results, pooling regional neighborhood counts can increase the speed of detection in comparison with individual regional counts, when the shift, the deviation from the mean, occurs in several units that make up the regional neighborhood (this was previously emphasized by Raubertas [7]). However, pooling may result in a delay in detection if the shift occurs in only one or two regions that make up a neighborhood because the effect of the shift will be diluted. Additionally, pooling may result in an increase in false alarms in regions sharing a common boundary. Another issue related to the utilization of the CUSUM statistic is in identifying the time at which the outbreak ends. The CUSUM statistic may not decrease quickly after an outbreak has ended, leading to incorrect decisions when the outbreak is over. This issue is addressed by Gandy and Lau [26].

The proposed methodology is applicable for the detection of foodborne outbreaks, a class of outbreaks with no trend or seasonality. A future research direction is to extend the method to a broader class of settings, encompassing outbreaks with trend and seasonality. Additionally, we have assumed spatial and temporal independence of the observed counts (although the pooled CUSUM statistics are correlated). It is certainly possible that the counts are dependent, even after adjusting for a nonstationary mean structure. This has been addressed by Rogerson and Yamada [8] among others. We are actively working on extending the proposed methodology to the setting where counts are dependent in time and/or space.

Acknowledgements

The R package 'fdrtool' [27] was used to calculate the q-values for the simulations and the case study. We would like to thank several anonymous referees and an associate editor for their very helpful suggestions toward improving this manuscript.

References

1. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society A* 2012; **175**(1):49–82.
2. Woodall WH. The use of control charts in health-care and public health surveillance. *Journal of Quality Technology* 2006; **38**:89–104.
3. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. *Journal of Royal Statistical Society A* 2003; **166**:5–21.
4. Tsui KL, Chiu W, Gierlich P, Goldsman D, Liu X, Maschek T. A review of healthcare, public health, and syndromic surveillance. *Quality Engineering* 2008; **20**(4):435–450.
5. Page ES. Continuous inspection schemes. *Biometrika* 1954; **41**:100–115.
6. Lucas JM. Counted data CUSUM's. *Technometrics* 1985; **27**:129–144.
7. Raubertas RF. An analysis of disease surveillance data that uses the geographical locations of the reporting units. *Statistics in Medicine* 1989; **8**:267–271.
8. Rogerson PA, Yamada I. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine* 2004; **23**:2195–2214.
9. Li Y, Tsung F. Multiple attribute control charts with false discovery rate control. *Quality and Reliability Engineering International* 2012; **28**(6):857–871.
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 1995; **57**:289–300.
11. Benjamini Y, Yekutieli D. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* 2005; **469**:71–81.
12. Lee SH, Park JH, Jun C. An exponentially weighted moving average chart controlling false discovery rate. *Journal of Statistical Computation and Simulation* 2014; **84**(8):1830–1840.
13. Li Z, Qui P, Chatterjee S, Wang Z. Using p values to design statistical process control charts. *Statistical Papers* 2013; **54**(2):523–539.
14. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA* 2003; **100**:9440–9445.
15. Robertson C, Nelson TA, MacNab YC, Lawson AB. Review of methods for space-time disease surveillance. *Spatial and Spatio-temporal Epidemiology* 2010; **1**:105–116.
16. Siegmund D. Sequential Analysis: Tests and Confidence Intervals. Springer-Verlag: New York, 1985.
17. Moustakides GV. Optimal stopping times for detecting changes in distributions. *Annals of Statistics* 1986; **14**(4):1379–1387.
18. Rogerson P. Monitoring point patterns for the development of space time clusters. *Journal of the Royal Statistical Society A* 2001; **164**(1):87–96.

19. Purdy GG, Richards SC, Woodall WH. Surveillance of nonhomogeneous Poisson processes. *Technometrics* 2015; **57** (3):388–394.
20. Efron B, Storey JD, Tibshirani R. Technical report 2001–217 (Stanford University, Palo Alto, CA), 2001.
21. Frisén M. Evaluations of methods for statistical surveillance. *Statistics in Medicine* 1992; **11**:1489–1502.
22. Jiang W, Shu L, Zhao H, Tsui KL. CUSUM procedures for health care surveillance. *Quality and Reliability Engineering International* 2013; **29**(6):883–897.
23. The data are queried from the Survstat@RKI database of the German Robert Koch Institute: http://www3.rki.de/SurvStat/. Accessed March 16, 2015.
24. Hawkins DM, Olwell DH. Cumulative Sum Charts and Charting for Quality Improvement. Springer: New York, NY, 1998; 119–120.
25. Bayer C, Bernard H, Prager R, Rabsch W, Hiller P, Malorny B, Pfefferkorn B, Frank C, de Jong A, Friesema I, Start K, Rosner BM. An outbreak of *Salmonella* Newport associated with mung bean sprouts in Germany and the Netherlands, October to November 2011. *Eurosurveillance* 2014; **19**(1):1–9.
26. Gandy A, Lau FD. Non-restarting cumulative sum charts and control of the false discovery rate. *Biometrika* 2013; **100**:261–268.
27. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* 2008; **24**(12):1461–1462.