Research Article

Trigonometric Regression for Analysis of Public Health Surveillance Data

Steven E. Rigdon,[1] George Turabelidze,[2] and Ehsan Jahanpour[2]

[1] *Department of Biostatistics, College for Public Health and Social Justice, Saint Louis University, 3545 Lafayette Avenue, St. Louis, MO 63104, USA*

[2] *Missouri Department of Health and Senior Services, 220 S. Jefferson Avenue, St. Louis, MO 63103, USA*

Correspondence should be addressed to Steven E. Rigdon; srigdon@slu.edu

Statistical challenges in monitoring modern biosurveillance data are well described in the literature. Even though assumptions of normality, independence, and stationarity are typically violated in the biosurveillance data, statistical process control (SPC) charts adopted from industry have been widely used in public health for communicable disease monitoring. But, blind usage of SPC charts in public health that ignores the characteristics of disease surveillance data may result in poor detection of disease outbreaks and/or excessive false-positive alarms. Thus, improved biosurveillance systems are clearly needed, and participation of statisticians knowledgeable in SPC alongside epidemiologists in the design and evaluation of such systems can be more productive. We describe and study a method for monitoring reportable disease counts using a Poisson distribution whose mean is allowed to vary depending on the week of the year. The seasonality is modeled by a trigonometric function whose parameters can be estimated by some baseline set of data. We study the ability of such a model to detect an outbreak. Specifically, we estimate the probability of detection (POD), the average number of weeks to signal given that a signal has occurred (conditional expected delay, or CED), and the false-positive rate (FPR, the average number of false-alarms per year).

1. Introduction

Homeland Security Presidential Directive 21 of October 18, 2007, defined biosurveillance as "... the process of active data-gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health—whether infectious, toxic, metabolic, or otherwise, and regardless of intentional or natural origin—in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity...." This suggests two distinct goals: situational awareness and early event detection. While situational awareness is certainly an important characteristic of a disease surveillance system, we focus on the problem of early event detection (EED).

Many diseases have an occurrence rate that is periodic throughout the year. Often a disease will peak in the summer and have a low occurrence rate in the winter, or vice versa.

For example, *E. coli O157:H7* infections tend to peak in the summer and pertussis tends to peak in the winter. There are some diseases, such as tuberculosis, that have a nearly constant occurrence rate throughout the year. A surveillance system for diseases whose occurrence rate is seasonal cannot have a constant control limit since there would be frequent signal limits in the peak season and practically no signals in the off-season. In this paper, we propose and study the use of a seasonal model for the mean, or expected, disease count. Data are typically collected weekly and the Poisson distribution is a reasonable model for these disease counts. Figure 1 illustrates the seasonality in the reported cases of pertussis in the state of Missouri from 2002 through 2011. The rate is least in late winter, early spring, and highest in late fall and throughout most of the winter. The control limits, which are shown in red in this figure, will be discussed in the next section.

The problem of disease surveillance or biosurveillance has been addressed in a number of books in recent years.
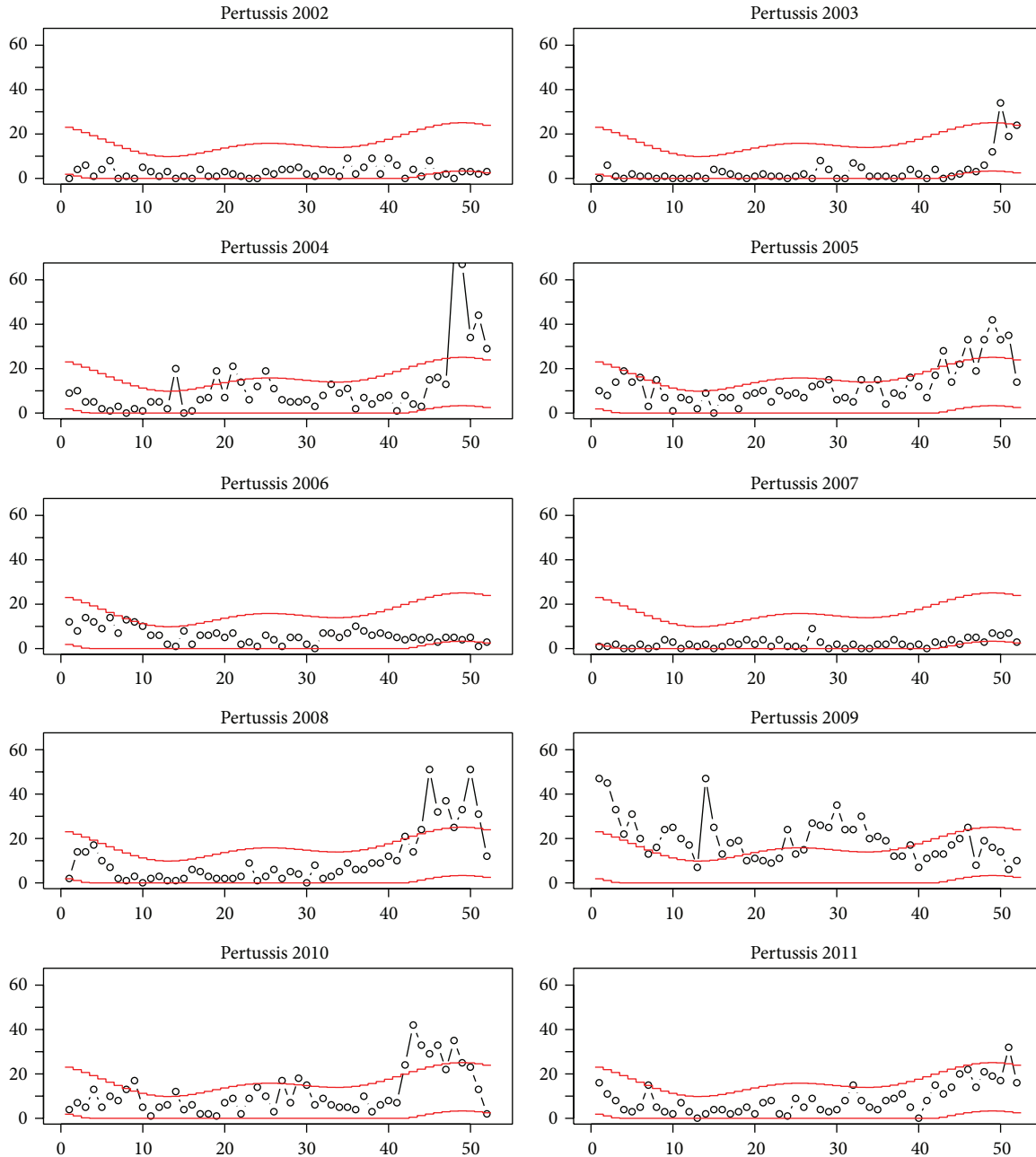
Figure 1: Weekly Pertussis cases, 2002–2011, Missouri, USA.

The edited books by Kass-Hout and Zhang [1], Lawson and Kleinman [2], Lombardo and Buckeridge [3], M'ikanatha et al. [4], Wagner et al. [5], Wilson et al. [6], and Zeng et al. [7] cover a wide spectrum of techniques for monitoring disease. Recently, the books by Chen et al. [8] and Fricker [9] have systematically studied the problem of disease surveillance.

Disease surveillance shares many characteristics with quality control. In quality control or quality surveillance, control charts are used to monitor the quality of a manufacturing process. When a special event occurs, causing the output of the process to abruptly change, the control chart should raise a signal for workers to investigate. Thus, a control chart is used in industry to identify when the output differs from what would be expected by chance. In a similar fashion, disease surveillance is supposed to detect when the occurrence of a disease exceeds what would be expected by chance because this aberration in data may indicate a probable disease outbreak. Thus, the primary tool for disease surveillance could be a control chart, similar to what is used in industry. This technique is called statistical process control (SPC).

There are, however, differences between biosurveillance and SPC, and SPC methods must be modified before using them for biosurveillance analysis. While quality control has

the single objective of detecting process changes, biosurveillance has the dual goals of early event detection and situational awareness. In quality control, no effort is placed on estimating the current state of the system, since in the absence of a signal, the process mean is assumed unchanged. In quality control, when a chart does raise an out-of-control signal, the process is stopped and a search is made for the cause. When the process resumes, the control charts are reset when monitoring begins after a signal. By contrast, in biosurveillance, there is no "stopping of the process" and as a result, process monitoring continues and the monitoring statistics are never reset. Even without intervention, disease outbreaks that cause out-of-control signals are transient, and disease incidence returns to its original rate when the outbreak is over. Because of this, the metric of average run length (ARL), which presumes a prolonged constant step shift, is inappropriate for comparing methods of surveillance. Montgomery [10] gave a comprehensive account of methods for statistical process control, including charts for counts, and Woodall [11] gave a review of methods for discrete charts and provided a thorough review of literature up to that point in time. The Poisson distribution, which plays a central role in disease surveillance is monitored using the c-chart or the u-chart. Borror et al. [12] proposed use of an exponentially weighted moving average control chart for Poisson data which has better run length properties for small shifts, and Jonsson [13] proposed use of the CUSUM. Farrington et al. [14] developed methods for detecting the occurrence of outbreaks using a scanning system along with a simple regression algorithm. Parker [15] specifically applied the Poisson distribution to problems of disease surveillance. Recently, Noufaily et al. [16], Unkel et al. [17], and Enki et al. [18] developed and applied methods for Poisson counts and quasi-Poisson counts. Shmueli and Burkom [19] and Fricker [20] discuss some of the methodological issues in disease surveillance. Serfling [21] was the first to suggest modeling seasonality in disease rates when he studied the spread of influenza.

In this paper, we describe a simple method for early outbreak detection, DESTEM (Disease Electronic Surveillance with Trigonometric Models), and we apply it to the reportable communicable diseases data in Missouri, USA.

2. The Trigonometric Model

We assume that the disease count in week t follows a Poisson distribution with mean μ_t where μ_t varies by week according to a first-order trigonometric model

$$\mu_t = a_0 + b_1 \sin\left(\frac{2\pi t}{52}\right) + c_1 \cos\left(\frac{2\pi t}{52}\right) \tag{1}$$

or a second-order trigonometric model

$$\begin{aligned} \mu_t = a_0 &+ b_1 \sin\left(\frac{2\pi t}{52}\right) + b_2 \cos\left(\frac{2\pi t}{52}\right) \\ &+ c_1 \sin\left(\frac{4\pi t}{52}\right) + c_2 \cos\left(\frac{4\pi t}{52}\right). \end{aligned} \tag{2}$$
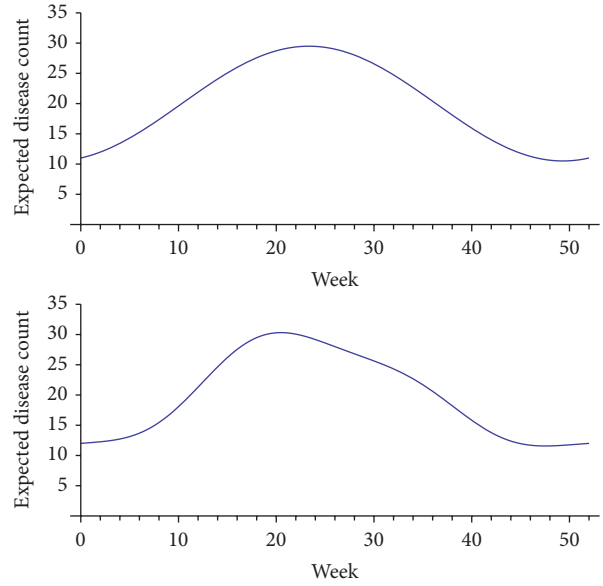

FIGURE 2: First-order trigonometric model (top) and second-order trigonometric model (bottom). The second-order model can be used in situations where the disease "lingers" after peaking.

Of course, a higher-order model, such as

$$\mu_t = a_0 + \sum_{k=1}^{p} \left(b_k \sin\left(\frac{2k\pi t}{52}\right) + c_k \cos\left(\frac{2k\pi t}{52}\right) \right) \tag{3}$$

could be used, but the first-order and second-order models are sufficient to model most diseases. Also, for diseases that exhibit no seasonality, a zeroth order model, $\mu_t = a_0$ can be applied. A first-order trigonometric model can model simple seasonality, where the expected counts follow a sinusoidal curve. A second-order trigonometric model can model situations where a disease "lingers" for a while once the disease peaks. Figure 2 shows the added flexibility in a second-order model.

For the Poisson distribution, the variance equals the mean, or, equivalently, the standard deviation equals the square root of the mean. As mentioned in the previous section, the main reason for monitoring diseases is to detect when an outbreak occurs. In order to do this, either in disease surveillance or quality surveillance, it is common to put an upper limit (or both an upper and a lower limit) and when observed points are outside the limits, it is inferred that a change in the process has occurred and an investigation should be done into the nature and cause. In this way, quality surveillance and disease surveillance are similar. Usually, however, in disease surveillance we are interested in whether there has been an increase in the underlying disease rate μ_t; this suggests using an upper control limit only. Thus, when an observed count exceeds the upper limit, an outbreak is inferred. The most common approach is to put limits three standard deviations above the mean (also, three standard deviations below the mean in the case of quality surveillance). Since the standard deviation of the Poisson is equal to the

square root of the mean, we have the following result for the upper control limit:

$$UCL = \mu_t + L\sqrt{\mu_t}. \tag{4}$$

The most common choice for L is $L = 3$, since for most distributions nearly all of the probability lies with three standard deviations of the mean. For a discrete distribution like the Poisson, this probability is quite variable since the output is constrained to be an integer. For example, if the mean is $\mu_t = 6.4$, then the standard deviation is 2.5298 and the upper control limit is 13.99, leaving a probability of

$$P(X > 13.99) = P(X \geq 14) = \sum_{k=14}^{\infty} P(X = k) \tag{5}$$
$$= 0.00625$$

of exceeding the upper control limit. (Here X indicates the random count for the week.) On the other hand, if the mean is just slightly higher, say $\mu_t = 6.5$, then the upper control limit is 14.63, leaving a probability of

$$P(X > 14.63) = P(X \geq 15) = \sum_{k=15}^{\infty} P(X = k) = 0.00339 \tag{6}$$

of exceeding the upper control limit. While these are both small numbers, their reciprocals, $1/0.00625 = 160.0$ and $1/0.00339 = 294.7$, which have an interpretation as the expected number of time periods for a signal (assuming that the means stays constant), are quite different.

Thus, a small change in the mean, from 6.4 to 6.5, caused a significant change in the probability of exceedance, with one being almost double the other.

The trigonometric models in (1) and (2) assume that the disease pattern remains constant from year to year. This assumption can be relaxed by incorporating one or more terms that depend on time t; for example, a simple linear trend can be accounted for by modifying the models to be of the form

$$\mu_t = a_0 + a_1 t + b_1 \sin\left(\frac{2\pi t}{52}\right) + b_2 \cos\left(\frac{2\pi t}{52}\right) \tag{7}$$

or

$$\mu_t = a_0 + a_1 t + b_1 \sin\left(\frac{2\pi t}{52}\right) + b_2 \cos\left(\frac{2\pi t}{52}\right)$$
$$+ c_1 \sin\left(\frac{4\pi t}{52}\right) + c_2 \cos\left(\frac{4\pi t}{52}\right). \tag{8}$$

For some diseases, such as tick-borne ehrlichiosis, there is a strong increasing trend across time that must be taken into account, and for other diseases, such as giardiasis, there is a decreasing trend. We have found that for most such diseases, a linear trend as in (7) or (8) is adequate.

It is possible that the predicted means near the trough, the lowest point on the curve, are negative if the models in (1) through (8) are used. This phenomenon actually occurs for some diseases. There are two approaches to dealing with this problem. One is to truncate the expected count at 0 and make a single case cause a signal. This is desirable for some diseases, such as anthrax, where a single occurrence of a disease is evidence of an outbreak. The second approach is to assume that the expected count is actually the exponential of the value shown in (1) to (8). For example, the second-order model with a linear trend would have mean function

$$\mu_t = \exp\left(a_0 + a_1 t + b_1 \sin\left(\frac{2\pi t}{52}\right) + b_2 \cos\left(\frac{2\pi t}{52}\right)\right.$$
$$\left. + c_1 \sin\left(\frac{4\pi t}{52}\right) + c_2 \cos\left(\frac{4\pi t}{52}\right)\right). \tag{9}$$

For only a few of the diseases that are monitored does the estimated mean function dip below zero, so we have found that applying one of the formulas (1) through (8) is sufficient. Even in these cases the estimated mean function will usually dip below zero for just a few weeks during the year, and in these cases a single case is sufficient to raise a signal.

Figure 1 shows the pertussis cases in Missouri from 2002 through 2011 with the upper control limits, calculated by $UCL = \mu_t + L\sqrt{\mu_t}$, where μ_t is the second-order model from (2). For pertussis, the estimated equation for μ_t is

$$\mu_t = 8.0592 - 2.2697 \sin\left(\frac{2\pi t}{52}\right) + 2.5403 \cos\left(\frac{2\pi t}{52}\right)$$
$$- 1.3465 \sin\left(\frac{4\pi t}{52}\right) + 1.4236 \cos\left(\frac{4\pi t}{52}\right). \tag{10}$$

Figure 3 shows the more recent case of *E. coli O157:H7* infections for the year 2013. For *E. coli O157:H7*, a first-order model is adequate. Here the estimated trigonometric curve, from (1), is

$$\mu_t = 1.5420 - 0.9203 \sin\left(\frac{2\pi t}{52}\right) - 0.9363 \cos\left(\frac{2\pi t}{52}\right). \tag{11}$$

Calculations for the method and the resulting plots were done using the package R [22]. The regressions were done using the linear model function (lm) in R. Figure 3 shows the system, developed in R [22], that is used to do the required calculations and display the graphs.

Typically, 50 to 60 diseases (of the 152 reportable diseases) are monitored each week. A summary is prepared for all of the monitored diseases. The report gives the current week count, the year to date count, the count from the same week of the previous year, a flag denoting the status, and the upper control limit for the next week. The flag is red if the point is above the upper control limit, and blue otherwise. The next week's upper control limit is given so that a signal can be raised at the earliest possible moment. For example, if next week's UCL is 2.4 and three cases have been observed by the middle of the week, then a signal can be raised at that point. Table 1 shows part of the weekly summary based on the DESTEM procedure of the status of reportable diseases for the state of Missouri in Week 17 of 2014, the week ending April 26, 2014.

TABLE 1: A portion of the weekly report for the week ending April 26, 2014, showing the status of just the first eleven diseases. The red flag (circled) indicates a point above the upper control limit. Altogether, there were three "red flags" for this week, with the other two not shown in this table.

Disease	Week 17 count	YTD count	Previous YTD count	Flag	Next week upper control limit
Anaplasma phagocytophilum	1	3	0	🚩	1.1
Botulism infant	0	2	1		0
Brucellosis	0	0	1		0
Campylobacteriosis	8	99	163		24.1
Coccidioidomycosis	0	5	2		1.3
Creutzfeldt-Jakob disease (CJD)	0	1	5		1
Cryptosporidiosis	1	36	50		6.2
Cyclosporiasis	0	0	0		0
Dengue fever	0	2	0		0
E. coli Shiga toxin positive	0	24	21		5.4
E. coli O157	0	11	19		4.8

Missouri Department of Health and Senior Services.
DESTEM report for Week 17, week ending, April 26, 2014.
State.
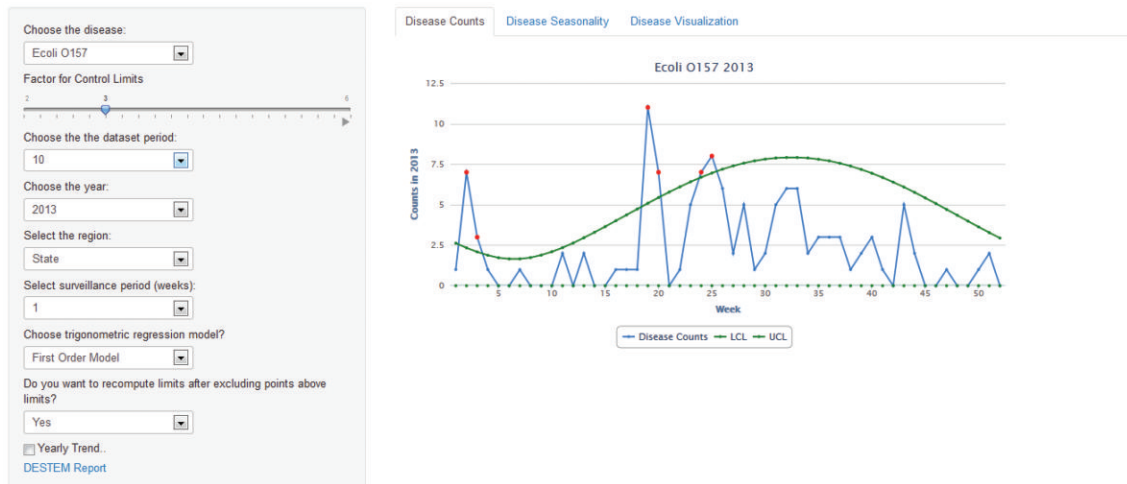
FIGURE 3: *E. coli O157:H7* infections for the year 2013 with the upper control limits computed using the first-order trigonometric model in (1).

3. A Simulation Study of DESTEM Performance

In order to test the method's ability to detect aberrations indicative of a possible disease outbreak, a Monte-Carlo simulation is used to model a number of processes with an induced outbreak. The typical outbreak is one where the incidence increases linearly for c time periods, reaching a maximum of b additional cases, then, decreasing linearly for c time periods when the number of additional cases is back to zero. Figure 4 illustrates this type of outbreak and how it is applied to the simulated data. Following the suggestions of Fraker et al. [23], we use simulation to estimate the following quantities.

(1) The probability of outbreak detection (POD): this is the probability that the outbreak is detected while it is still occurring. A small outbreak for a short period of time will have a smaller probability of being detected than a large outbreak for a longer period of time.

(2) Conditional expected delay (CED): this is the expected number of time periods until a signal is reached conditioned on there being a signal.

(3) False-positive rate (FPR): this is the average number of false-alarms that are signaled per year.

The usual metric of average run length (ARL), which is ubiquitous in the process control literature and widely used in disease surveillance, is inappropriate for the application

Figure 4: Illustration of typical outbreak and its effect on the expected disease count. The bottom figure shows the outbreak times used in the simulations.

discussed here. Use of the ARL presupposes a prolonged step shift in the variable being monitored, which is unrealistic for most diseases. More reasonable is an outbreak for which the number of cases increases for a while and then, even in the absence of an intervention, decreases back to the background noise.

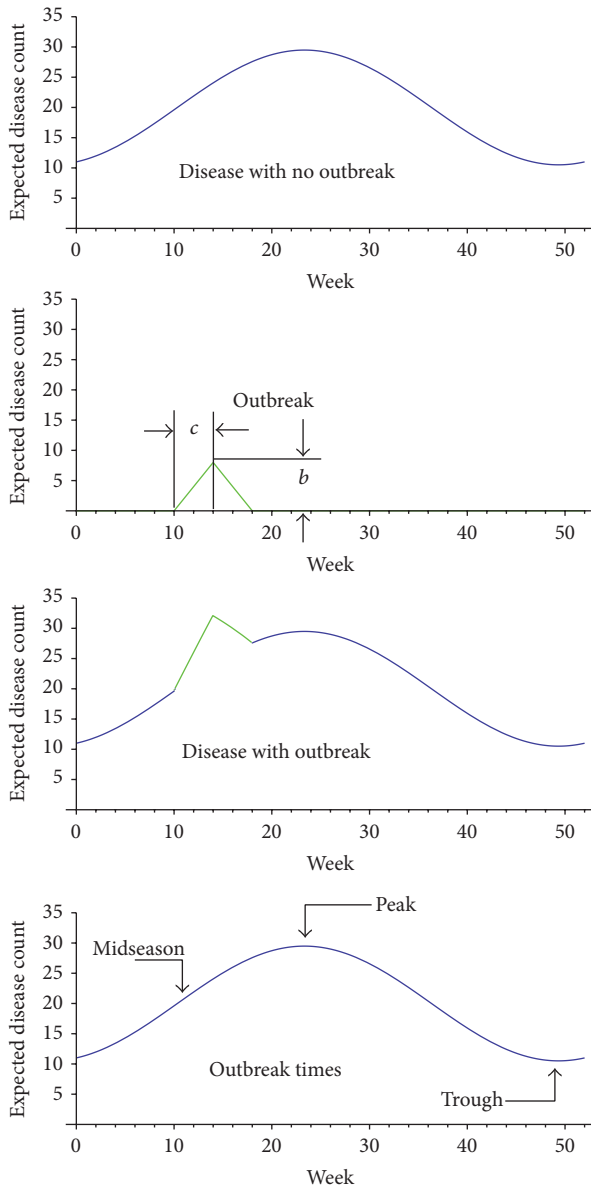We used a 10-year database of Missouri's surveillance of two epidemiologically distinct diseases: *E. coli O157:H7* infections and pertussis. For pertussis, we selected a second-order model and for *E. coli O157:H7* infections we selected a first-order model. These choices were based on the epidemiological curve of those diseases across the ten-year data base. In order to obtain baseline data, the weeks where an outbreak was obviously occurring were removed from the

data base. Because most outbreaks usually occur within a defined geographic region rather than in the entire state at once, we tested DESTEM on both, the regional and the state level data.

3.1. First-Order Trigonometric Model Simulation for E. coli O157:H7. The experimental factors included outbreak start time, maximum outbreak size, and outbreak duration. One outbreak per year was considered in the simulation which could start at trough, midrange, or peak time of the yearly disease cycle. The outbreak duration for *E. coli O157:H7* was 2 (short), 3 (medium), or 4 (long) weeks. Two sets of outbreaks, small and large, were imposed in the simulation. The magnitude of small and large outbreaks varied by the outbreak start time. The outbreak sizes considered for *E. coli O157:H7* were 1 and 2 additional cases per week during the trough time, and 2 and 4 additional cases per week during the disease midrange and peak occurrence. In total, 18 scenarios were designed to assess the performance of DESTEM in detecting disease outbreaks. Figure 5 illustrates four years of a Monte-Carlo simulation using the parameters of scenario 15 (size = 2, duration = 4, signature = $(1, 2, 2, 1)$, total cases = 6, occurring at peak) along with real *E. coli O157:H7* counts in Missouri's Eastern region during 2010–2013. The signature of an outbreak gives the number of additional expected cases for each week during the outbreak; for example, a signature of $(1, 3, 1)$ would indicate a three-week outbreak with one additional case the first week, three the second week, and one the third week. The blue line shows fitted first-order trigonometric curve used to model the average weekly counts. The circles are the real *E. coli O157:H7* cases, and the triangles are simulated counts. The simulated outbreaks occurred at disease peak (week 32) and are shown in red triangles. The outbreak duration was 4 weeks and the size was 2 additional cases per week.

Each scenario was simulated 10,000 times and the average POD, CED, and FPR were computed. The simulation results, shown in Table 2, indicate that the probability of detecting an outbreak at the regional level was higher and the detection was faster (CED values were smaller) when compared to the state level data. The false-positive signal rate was higher for the regional data analysis; however, the overall false-positive rate was small for all scenarios. On average, one false-alarm was produced by DESTEM in 1.91 years at the state level, and one false-alarm was produced in 1.26 years at the regional level for *E. coli O157:H7* infections.

3.2. Second-Order Trigonometric Model Simulation for Pertussis. Since pertussis cases occur more frequently than *E. coli O157:H7* cases and the outbreaks are prolonged, the experimental parameters were changed. One outbreak per year was considered at trough, midrange, or peak of yearly occurrence and could last for 4 (short), 5 (medium), or 6 (long) weeks. The outbreak size during the pertussis trough time (Week 10) is assumed to be equal to either 5 or 10 additional cases per week, and during midrange and peak time (weeks 25 and 46, resp.) the imposed outbreak is considered to be either 10 or 20 additional cases per week. Figure 6 shows the actual pertussis

TABLE 2: Simulation results for *E. coli O157:H7* infection outbreaks assuming Poisson model.

	E. coli-like outbreak			Probability of detection[1]		Conditional expected delay[2]		False-positive rate[3]	
Size (*b*)	Duration (2*c*)	Signature[4]	Time	Region	State	Region	State	Region	State
1	2	(1, 1): 2	Trough	1.000	0.373	1.000	1.433	0.797	0.531
1	3	(1, 1, 1): 3	Trough	1.000	0.516	1.000	1.878	0.832	0.482
1	4	(1, 1, 1, 1): 4	Trough	1.000	0.637	1.000	2.282	0.811	0.459
2	2	(2, 2): 4	Trough	1.000	1.000	1.000	1.000	0.825	0.521
2	3	(1, 2, 1): 4	Trough	1.000	1.000	1.000	1.798	0.799	0.494
2	4	(1, 2, 2, 1): 6	Trough	1.000	1.000	1.000	1.796	0.803	0.462
2	2	(2, 2): 4	Mid	0.545	0.142	1.451	1.586	0.835	0.553
2	3	(1, 2, 1): 4	Mid	0.421	0.133	2.007	2.107	0.791	0.546
2	4	(1, 2, 2, 1): 6	Mid	0.653	0.207	2.417	2.476	0.762	0.537
4	2	(4, 4): 8	Mid	1.000	0.708	1.000	1.393	0.826	0.558
4	3	(2, 4, 2): 8	Mid	1.000	0.573	1.701	1.979	0.797	0.534
4	4	(2, 4, 4, 2): 12	Mid	1.000	0.789	1.694	2.281	0.784	0.536
2	2	(2, 2): 4	Peak	0.457	0.136	1.432	1.453	0.814	0.548
2	3	(1, 2, 1): 4	Peak	0.371	0.119	1.954	2.004	0.783	0.540
2	4	(1, 2, 2, 1): 6	Peak	0.543	0.183	2.358	2.414	0.740	0.535
4	2	(4, 4): 8	Peak	1.000	0.544	1.000	1.410	0.791	0.529
4	3	(2, 4, 2): 8	Peak	1.000	0.405	1.739	1.931	0.773	0.532
4	4	(2, 4, 4, 2): 12	Peak	1.000	0.596	1.747	2.295	0.757	0.536

[1]Probability of detection (POD) per "outbreak."
[2]Conditional expected delay (CED) per "detected outbreaks."
[3]False-positive rate in terms of false-positive signals per year.
[4]The signature is the number of cases in each week during the outbreak, along with the total number of cases.

FIGURE 5: Simulated (triangles) and real (circles) *E. coli O157:H7* infection counts, Missouri, Eastern Region, 2010–2013. Simulated outbreaks are the red triangles. The fitted trigonometric model is the blue curve.

data (circles) along with the simulated data (triangles) using the parameters of scenario 15 in Table 3 (size = 10, duration = 6, signature = (4, 7, 10, 10, 7, 4), with a total of 42 additional cases, occurring at peak). Although we found a seasonal component to the counts of pertussis cases, others such as de Greeff et al. [24] have found an even stronger seasonal component. We then ran the simulation 10,000 times for each of the 18 scenarios (Table 3). DESTEM performed well in detecting pertussis outbreaks accurately and timely. The false-positive signal rate was small: roughly one false-alarm every 5.35 years at the state level and one false-alarm every

3.89 years at the regional level. The probability of detection was nearly 1 for many outbreak scenarios and the conditional expected delay was usually about 2 or under.

3.3. Performance of DESTEM When There Is Overdispersion.
The Poisson distribution is the simplest model for counts, and it can be an appropriate model for counts in a number of situations. The Poisson distribution, being a one-parameter distribution, has the property that the mean and the variance are equal. In many applications, the variance of the data exceeds the mean, making the Poisson model inappropriate.

TABLE 3: Simulation results for pertussis outbreaks assuming outbreak model.

	Outbreak			Probability of detection[1]		Conditional expected delay[2]		False-positive rate[3]	
Size (b)	Duration ($2c$)	Signature[4]	Time	Region	State	Region	State	Region	State
5	4	(3, 5, 5, 3): 16	Trough	0.863	0.518	1.903	1.228	0.268	0.187
5	5	(2, 4, 5, 4, 2): 17	Trough	0.758	0.429	2.097	1.237	0.249	0.181
5	6	(1, 3, 5, 5, 3, 1): 18	Trough	0.908	0.583	2.700	1.881	0.255	0.174
10	4	(5, 10, 10, 5): 30	Trough	1.000	0.999	1.371	1.790	0.259	0.187
10	5	(4, 7, 10, 7, 4): 32	Trough	1.000	0.998	1.836	2.255	0.244	0.182
10	6	(4, 7, 10, 10, 7, 4): 42	Trough	1.000	1.000	1.839	2.267	0.255	0.186
10	4	(5, 10, 10, 5): 30	Mid	1.000	0.976	1.577	1.922	0.243	0.192
10	5	(4, 7, 10, 7, 4): 32	Mid	1.000	0.939	2.039	2.391	0.252	0.184
10	6	(4, 7, 10, 10, 7, 4): 42	Mid	1.000	0.987	2.038	2.580	0.244	0.185
20	4	(10, 20, 20, 10): 60	Mid	1.000	1.000	1.000	1.172	0.257	0.186
20	5	(7, 14, 20, 14, 7): 62	Mid	1.000	1.000	1.153	1.616	0.260	0.177
20	6	(7, 14, 20, 20, 14, 7): 82	Mid	1.000	1.000	1.158	1.607	0.244	0.185
10	4	(5, 10, 10, 5): 30	Peak	0.946	0.629	1.965	1.437	0.272	0.195
10	5	(4, 7, 10, 7, 4): 32	Peak	0.881	0.524	2.342	1.504	0.261	0.194
10	6	(4, 7, 10, 10, 7, 4): 42	Peak	0.970	0.704	2.723	2.262	0.263	0.194
20	4	(10, 20, 20, 10): 60	Peak	1.000	1.000	1.276	1.582	0.268	0.197
20	5	(7, 14, 20, 14, 7): 62	Peak	1.000	0.999	1.715	2.119	0.269	0.190
20	6	(7, 14, 20, 20, 14, 7): 82	Peak	1.000	1.000	1.714	2.104	0.264	0.190

[1] Probability of detection (POD) per "outbreak."
[2] Conditional expected delay (CED) per "detected outbreaks."
[3] False-positive rate in terms of false-positive signals per year.
[4] The signature is the number of cases in each week during the outbreak, along with the total number of cases.
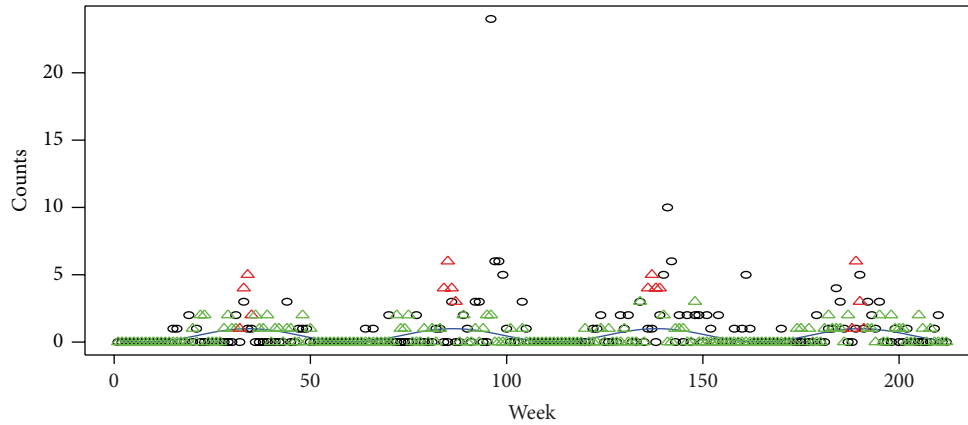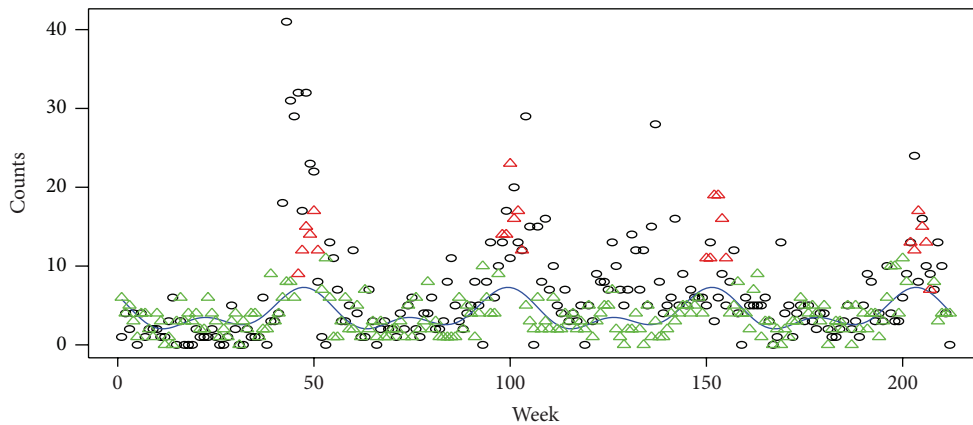

FIGURE 6: Simulated (triangles) and real (circles) Pertussis case counts, Missouri, Eastern region, 2010–2013. Simulated outbreaks are the red triangles. The fitted trigonometric model is the blue curve.

This concept is called *overdispersion*. As a remedy, the negative binomial distribution is often used as an alternative to the Poisson (Sparks et al. [25]). The negative binomial distribution is often thought of as the number of failures in a sequence of Bernoulli trials before obtaining the rth success. The probability mass function for the random variable Y with negative binomial distribution is

$$P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1 - p)^y, \quad y = 0, 1, 2, \ldots. \tag{12}$$

While the description given above requires the parameter r to be a positive integer, the above formula is a valid probability mass function for every $r > 0$, so long as the gamma distribution is used instead of factorials. Thus, for an arbitrary positive value r, the negative binomial has probability mass function

$$P(Y = y) = \frac{\Gamma(y + r)}{\Gamma(r)\, y!} p^r (1 - p)^{y - r}, \quad y = 0, 1, 2, \ldots. \tag{13}$$

FIGURE 7: Negative binomial fit to pertussis data (top), where points above UCL are shown in red. After these points were removed, the negative binomial was fit again, with the results shown in the bottom.
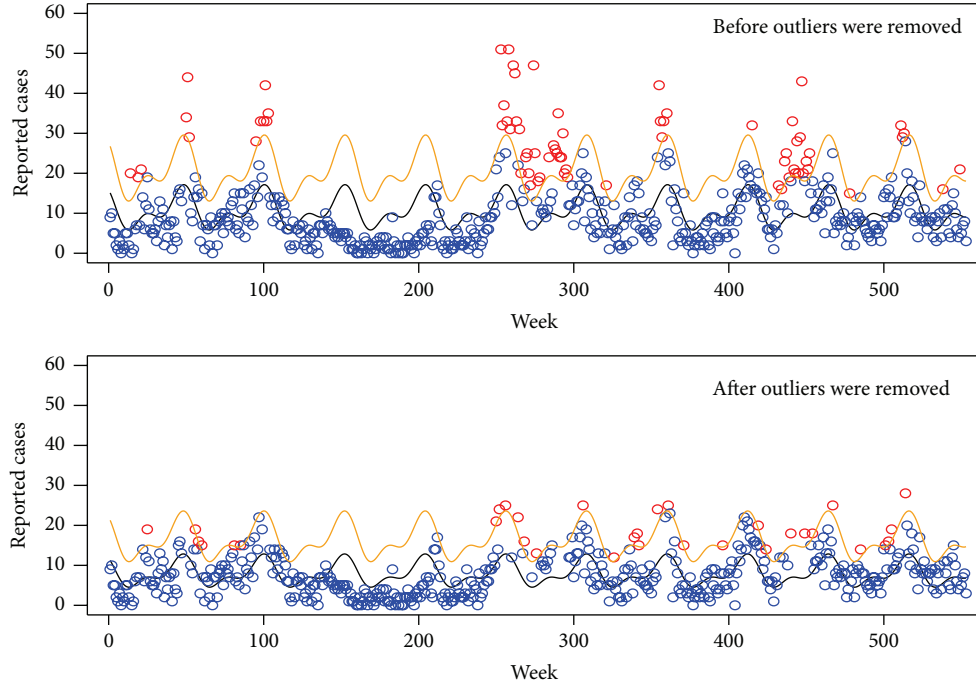
The mean and variance of the negative binomial distribution are $\mu_Y = r(1-p)/p$ and $\sigma^2 = r(1-p)/p^2$, so that the variance to mean ratio is $\sigma_Y^2/\mu_Y = 1/p > 1$. The negative binomial can be reparameterized so that the mean is equal to μ and the variance is $\sigma^2 = \mu + \theta\mu^2$. The parameter θ is called the over-dispersion parameter. A value of $\theta = 0$ leads to the variance being equal to the mean, and as $\theta \to 0$ the negative binomial approaches the Poisson.

We ran this negative binomial model on the pertussis data and found that the estimated mean function was

$$\mu_t = 10.7319 - 2.7483 \sin\left(\frac{2\pi t}{52}\right) + 2.9622 \cos\left(\frac{2\pi t}{52}\right)$$
$$- 1.5460 \sin\left(\frac{4\pi t}{52}\right) + 2.1789 \cos\left(\frac{4\pi t}{52}\right) \tag{14}$$

with an estimated overdispersion parameter of $\widehat{\theta} = 0.5392$. There were, however, a number of pertussis outbreaks during this period. These outbreaks seem to exaggerate the effect of overdispersion. If the points above the three standard deviation limits are removed, the estimates become

$$\mu_t = 8.0568 - 2.1514 \sin\left(\frac{2\pi t}{52}\right) + 2.3674 \cos\left(\frac{2\pi t}{52}\right)$$
$$- 1.1188 \sin\left(\frac{4\pi t}{52}\right) + 1.3768 \cos\left(\frac{4\pi t}{52}\right) \tag{15}$$

with an estimated overdispersion parameter of $\widehat{\theta} = 0.3215$. The estimated curves are shown in Figure 7. The top part of Figure 7 is the original data and the bottom part is the data with outliers removed. Although there are still some points

above the UCL in the bottom figure, these seem to be due to the higher variability and not to newly discovered outbreaks.

To see how the DESTEM algorithm works when the assumption of a Poisson distribution is violated, we ran several simulations under which the counts had a negative binomial distribution. For both *E. coli O157:H7*, where the counts averaged approximately 1.5 per week, and pertussis, where the counts averaged approximately 8 per week, we chose the overdispersion parameter to be $\theta = 0.1$, $\theta = 0.2$, or $\psi = 0.4$. This corresponds to a variance to mean ratio of

$$\frac{\sigma^2}{\mu} = \frac{\left(\mu + \theta\mu^2\right)}{\mu} = 1 + \theta\mu, \tag{16}$$

which is 1.15, 1.3, or 1.6 for *O157:H7* and 1.8, 2.6, and 4.2 for pertussis.

The simulations were performed much as in Tables 2 and 3. The observations were simulated according to a negative binomial coefficient with mean μ_t and overdispersion parameter θ. The Poisson distribution was then (incorrectly) assumed and the control limits were computed using three standard deviation limits. The results of the simulations are shown in Tables 4 and 5. Although the results vary a bit, the general conclusions from these tables are that the POD increases when there is overdispersion (desirable), but the FPR also increases (undesirable). These results are somewhat expected since the added variability will cause more signals, but when the process is stable, these will be false signals.

The DESTEM method here could have assumed the negative binomial distribution instead of the Poisson, and the upper control limit would have been a bit higher. This,

Table 4: Simulation results for *E. coli O157:H7* infection outbreaks assuming a negative binomial model for the entire state with overdispersion parameter $\theta = 0.1$, $\theta = 0.2$, or $\theta = 0.4$.

E. coli-like outbreak				Overdispersion $\theta = 0.1$			Overdispersion $\theta = 0.2$			Over dispersion $\theta = 0.4$		
Size (*b*)	Duration (*2c*)	Signature[4]	Time	POD[1]	CED[2]	FPR[3]	POD[1]	CED[2]	FPR[3]	POD[1]	CED[2]	FPR[3]
5	4	(1, 1): 2	Trough	1.000	1.000	0.932	1.000	1.000	1.067	1.000	1.000	1.321
5	5	(1, 1, 1): 3	Trough	1.000	1.984	0.939	1.000	1.985	1.069	1.000	1.985	1.306
5	6	(1, 1, 1, 1): 4	Trough	1.000	1.984	0.952	1.000	1.988	1.073	1.000	1.983	1.289
10	4	(2, 2): 4	Trough	1.000	1.000	0.910	1.000	1.000	1.060	1.000	1.000	1.305
10	5	(1, 2, 1): 4	Trough	1.000	1.000	0.930	1.000	1.000	1.060	1.000	1.000	1.303
10	6	(1, 2, 2, 1): 6	Trough	1.000	1.000	0.951	1.000	1.000	1.046	1.000	1.000	1.319
10	4	(2, 2): 4	Mid	0.533	1.446	0.934	0.523	1.444	1.070	0.518	1.431	1.271
10	5	(1, 2, 1): 4	Mid	0.426	2.009	0.922	0.429	2.000	1.057	0.432	2.009	1.266
10	6	(1, 2, 2, 1): 6	Mid	0.632	2.422	0.926	0.636	2.414	1.023	0.637	2.411	1.228
20	4	(4, 4): 8	Mid	1.000	1.000	0.938	1.000	1.000	1.068	1.000	1.000	1.299
20	5	(2, 4, 2): 8	Mid	1.000	1.701	0.929	1.000	1.704	1.037	1.000	1.709	1.268
20	6	(2, 4, 4, 2): 12	Mid	1.000	1.709	0.910	1.000	1.703	1.031	1.000	1.718	1.245
10	4	(2, 2): 4	Peak	0.452	1.422	0.901	0.441	1.433	1.041	0.447	1.428	1.219
10	5	(1, 2, 1): 4	Peak	0.385	1.936	0.906	0.386	1.942	1.012	0.414	1.921	1.203
10	6	(1, 2, 2, 1): 6	Peak	0.545	2.305	0.857	0.545	2.323	0.975	0.552	2.325	1.167
20	4	(4, 4): 8	Peak	1.000	1.000	0.917	1.000	1.000	1.030	1.000	1.000	1.244
20	5	(2, 4, 2): 8	Peak	1.000	1.745	0.890	1.000	1.748	1.003	1.000	1.747	1.208
20	6	(2, 4, 4, 2): 12	Peak	1.000	2.751	0.846	1.000	2.748	0.952	1.000	1.747	1.176

[1] Probability of detection (POD) per "outbreak."
[2] Conditional expected delay (CED) per "detected outbreaks."
[3] False-positive rate in terms of false-positive signals per year.
[4] The signature is the number of cases in each week during the outbreak, along with the total number of cases.

Table 5: Simulation results for *E. coli* O157:H7 infection outbreaks assuming a negative binomial model for the entire state with overdispersion parameter $\theta = 0.1$, $\theta = 0.2$, or $\theta = 0.4$.

		Pertussis-like outbreak		Overdispersion $\theta = 0.1$			Overdispersion $\theta = 0.2$			Overdispersion $\theta = 0.4$		
Size (*b*)	Duration (*2c*)	Signature[4]	Time	POD[1]	CED[2]	FPR[3]	POD[1]	CED[2]	FPR[3]	POD[1]	CED[2]	FPR[3]
5	4	(3, 5, 5, 3): 16	Trough	0.579	1.338	1.287	0.591	1.362	2.476	0.626	1.421	4.186
5	5	(2, 4, 5, 4, 2): 17	Trough	0.518	1.453	1.273	0.570	1.566	2.404	0.632	1.688	4.127
5	6	(1, 3, 5, 5, 3, 1): 18	Trough	0.652	2.024	1.255	0.687	2.099	2.428	0.723	2.127	4.069
10	4	(5, 10, 10, 5): 30	Trough	1.000	1.771	1.306	0.999	1.764	2.460	0.997	1.773	4.220
10	5	(4, 7, 10, 7, 4): 32	Trough	0.996	2.252	1.290	0.994	2.232	2.435	0.988	2.229	4.146
10	6	(4, 7, 10, 10, 7, 4): 42	Trough	1.000	2.244	1.255	1.000	2.259	2.393	0.999	2.256	4.078
10	4	(5, 10, 10, 5): 30	Mid	0.957	1.863	1.269	0.943	1.831	2.434	0.919	1.809	4.099
10	5	(4, 7, 10, 7, 4): 32	Mid	0.915	2.249	1.260	0.909	2.203	2.358	0.892	2.144	4.046
10	6	(4, 7, 10, 10, 7, 4): 42	Mid	0.977	2.519	1.229	0.970	2.470	2.339	0.958	2.433	3.972
20	4	(10, 20, 20, 10): 60	Mid	1.000	1.248	1.271	1.000	1.305	2.387	1.000	1.367	4.127
20	5	(7, 14, 20, 14, 7): 62	Mid	1.000	1.622	1.255	1.000	1.639	2.333	1.000	1.689	4.018
20	6	(7, 14, 20, 20, 14, 7): 82	Mid	1.000	1.620	1.221	1.000	1.650	2.344	1.000	1.678	3.950
10	4	(5, 10, 10, 5): 30	Peak	0.701	1.542	1.237	0.736	1.585	2.327	0.749	1.571	4.006
10	5	(4, 7, 10, 7, 4): 32	Peak	0.682	1.845	1.180	0.726	1.887	2.263	0.762	1.914	3.827
10	6	(4, 7, 10, 10, 7, 4): 42	Peak	0.799	2.375	1.152	0.834	2.398	2.196	0.856	2.339	3.734
20	4	(10, 20, 20, 10): 60	Peak	0.999	1.599	1.222	0.998	1.624	2.331	0.991	1.672	3.980
20	5	(7, 14, 20, 14, 7): 62	Peak	0.996	2.061	1.187	0.991	2.055	2.261	0.980	2.061	3.850
20	6	(7, 14, 20, 20, 14, 7): 82	peak	1.000	2.083	1.142	1.000	2.107	2.190	0.996	2.126	3.732

[1]Probability of detection (POD) per "outbreak."
[2]Conditional expected delay (CED) per "detected outbreaks."
[3]False-positive rate in terms of false-positive signals per year.
[4]The signature is the number of cases in each week during the outbreak, along with the total number of cases.

however, requires one additional important parameter to be estimated (the overdispersion parameter θ). Altogether, for the full model in (9) seven parameters would have to be estimated: $a_0, a_1, b_1, b_2, c_1, c_2$, and θ. It has been shown (Jensen et al. [26] and Champ et al. [27]) that very large sample sizes are often needed when the number of parameters is large. Misestimation of these parameters can lead to charts whose properties do not resemble those of the corresponding chart with assumed known parameters. For moderate sample sizes, the overdispersion parameter is estimated with a fairly large standard error. In addition, the existence of outbreaks makes the problem more difficult because we must remove the outbreak from the data set and reestimate the parameters. As a general rule, we have excluded observations that are above the upper control limit and refit the model based on the remaining observations. Of course, when these are removed the upper control limit drops and inevitably there are more points outside the limits. The process of removing outliers and recomputing limits can continue, but we have found that one or two iterations are sufficient. Whatever points remain outside can either be explained by overdispersion or still more outbreaks; in many cases, it is difficult to distinguish between these two possibilities.

4. Discussion and Conclusion

We developed a simple analytical methodology, DESTEM, that was able in this simulation study to timely and accurately detect aberrations in the reportable diseases weekly data. Aberrations in the usual communicable disease patterns may provide a warning sign of an outbreak, but detection of such aberrations still remains an important challenge in public health surveillance. Statistical simulation confirmed that DESTEM performs well under different scenarios routinely encountered by the epidemiologist at the Missouri Department of Health and Senior Services (DHSS).

Current widespread usage of syndromic surveillance systems by public health departments across the USA does not preclude necessity of traditional surveillance of reportable communicable diseases [28]. Syndromic surveillance systems, usually based on sophisticated algorithms, have major strengths, such as improved timeliness of data availability (often within hours), completeness of data, flexibility, reasonably good situational awareness, and so forth. But, unlike diagnostic data that underlies reportable communicable disease surveillance, syndromic data are an indirect indicator of a disease outbreak because it is based on the nonspecific symptom data rather than on the actual disease diagnosis. Systems such as Biosense [29], Essence [30], and others, are by contrast systems for collecting, archiving, and presenting surveillance data. The goals of syndromic surveillance are often short-term (such as for the Olympics, or a political convention) and look for omnibus signals in the data. By contrast, our reportable infections data are cyclical and we look for a deviation from this norm. See Chen et al. [8, chapter 2] for a summary of many of the popular surveillance systems.

The reportable diseases list includes a variety of conditions with very different incidence rates, incubation periods, seasonality, and urgency. Despite such a variety, DESTEM performed well when two epidemiologically distinct infection outbreaks, *E. coli O157:H7* and pertussis, were simulated. For the communicable disease epidemiologist, early detection of an outbreak is a priority. But, direct application to biosurveillance of the industrial SPC methods requiring sufficient data accumulation to detect change in the process mean could result in the delayed timeliness of detection, and therefore practical usefulness of surveillance system is diminished. Thus, the primary goal for DESTEM was set to meet a specific surveillance purpose, such as accurate and timely detection of disease outbreaks. The characteristic of DESTEM is that it does not directly detect a change in a disease trend; it rather detects an aberration that is different from a historical trend in any given week of the year. DESTEM analyzes whether the increased number of individual observations in any specific week of the year is unlikely to be happening by chance alone based on the modeled upper control limit that changes from week to week. Once the probability of such an occurrence by chance alone is determined to be too low, DESTEM produces a signal, prompting investigation by the epidemiologist regarding whether a true disease outbreak is occurring.

It is postulated that public health surveillance data is often overdispersed, and therefore negative binomial rather than Poisson distribution is more appropriate assumption (Sparks et al. [25]). Our simulation data showed that assuming negative binomial distribution for analysis may not be always advantageous: while the POD did increase, the FPR also went up. The problem of using estimated parameters instead of assumed known parameters arises when we must estimate the overdispersion. It is likely that both, Poisson and negative binomial distribution, can be reasonable assumptions when applied to surveillance data.

Simulation results were also consistent with a common epidemiological experience that all outbreaks start locally and should be analyzed using local data. DESTEM performed better when applied to the historical regional data rather than using state level data as a baseline for comparison. Thus, surveillance analysis methodology needs to be accurate, but it also needs to be applied to a more relevant and more precisely defined dataset.

We chose a relatively long 10-year historical baseline period for modeling because of the need for appropriate sample size and for better estimation of the variance. While producing statistically more stable estimates, longer historical periods could affect accuracy of those estimates due to systematic effects accumulating over the prolonged period of time. Naturally changing disease trends, new diagnostics, public health control measures, and changing attitudes of health care providers and population represent various systematic effects that may distort estimated baseline with a variable degree. Such a difficult tradeoff seemed to be justified based on the limited simulation analysis of DESTEM performance, but it will certainly need more vigorous assessment with the real public health surveillance data. In order to reduce possible systematic effects in the data, DESTEM renews the 10-year baseline dataset every year by discarding

the oldest year from the model while adding the most recent year's data.

In conclusion, DESTEM represents a promising tool for analysis of surveillance data of variety of reportable infectious diseases. The analytical reports based on this methodology are easy to understand not only for epidemiologists without advanced mathematical expertise, but for the general public as well.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] T. Kass-Hout and X. Zhang, Eds., *Biosurveillance: Methods and Case Studies*, Chapman & Hall/CRC, Boca Raton, FL, USA, 2011.

[2] A. B. Lawson and K. Kleinman, Eds., *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons, New York, NY, USA, 2005.

[3] J. S. Lombardo and D. L. Buckeridge, *Disease Surveillance: A Public Health Informatics Approach*, Wiley-Interscience, New York, NY, USA, 2007.

[4] N. M. M'ikanatha, R. Lynfield, C. A. van Beneden, and H. de Valk, Eds., *Infectious Disease Surveillance*, Wiley-Blackwell, New York, NY, USA, 2nd edition, 2013.

[5] M. W. Wagner, A. W. Moore, and R. M. Aryel, Eds., *Handbook of Biosurveillance*, Academic Press, Burlington, MA, USA, 2006.

[6] A. Wilson, G. Wilson, and D. H. Olwell, Eds., *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Springer, New York, NY, USA, 2006.

[7] D. Zeng, H. Chen, C. Castillo-Chavez, W. B. Lober, and M. Thurmond, Eds., *Infectious Disease Informatics and Biosurveillance*, Springer, New York, NY, USA, 2011.

[8] H. Chen, D. Zeng, and P. Yan, *Infectious Disease Informatics: Syndromic Surveillance for Public Health and Bio-Defense*, Springer, New York, NY, USA, 2010.

[9] R. D. Fricker, *Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance*, Cambridge University Press, 2013.

[10] D. C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, New York, NY, USA, 7th edition, 2013.

[11] W. H. Woodall, "Control charts based on attribute data: bibliography and review," *Journal of Quality Technology*, vol. 29, no. 2, pp. 172–183, 1997.

[12] C. M. Borror, C. W. Champ, and S. E. Rigdon, "Poisson EWMA control charts," *Journal of Quality Technology*, vol. 30, no. 4, pp. 352–361, 1998.

[13] R. Jonsson, "A CUSUM procedure for detection of outbreaks in Poisson distributed medical health events," Research Report, Statistical Research Unit, Department of Economics, University of Gothenburg, Gothenburg, Sweden, 2010, https://gupea.ub.gu.se/bitstream/2077/24352/1/gupea_2077_24352_1.pdf.

[14] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole, "A statistical algorithm for the early detection of outbreaks of infectious disease," *Journal of the Royal Statistical Society A*, vol. 159, no. 3, pp. 547–563, 1996.

[15] R. A. Parker, "Analysis of surveillance data with Poisson regression: a case study," *Statistics in Medicine*, vol. 8, no. 3, pp. 285–294, 1989.

[16] A. Noufaily, D. G. Enki, P. Farrington, P. Garthwaite, N. Andrews, and A. Charlett, "An improved algorithm for outbreak detection in multiple surveillance systems," *Statistics in Medicine*, vol. 32, no. 7, pp. 1206–1222, 2013.

[17] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: a review," *Journal of the Royal Statistical Society Series A*, vol. 175, no. 1, pp. 49–82, 2012.

[18] D. G. Enki, A. Noufaily, P. H. Garthwaite et al., "Automated biosurveillance data from England and Wales, 1991–2011," *Emerging Infectious Diseases*, vol. 19, no. 1, pp. 35–42, 2013.

[19] G. Shmueli and H. Burkom, "Statistical challenges facing early outbreak detection in biosurveillance," *Technometrics*, vol. 52, no. 1, pp. 39–51, 2010.

[20] J. Fricker, "Some methodological issues in biosurveillance," *Statistics in Medicine*, vol. 30, no. 5, pp. 403–415, 2011.

[21] R. E. Serfling, "Methods for current statistical analysis of excess pneumonia-influenza deaths," *Public Health Reports*, vol. 78, pp. 494–506, 1963.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, http://www.R-project.org/.

[23] S. E. Fraker, W. H. Woodall, and S. Mousavi, "Performance metrics for surveillance schemes," *Quality Engineering*, vol. 20, no. 4, pp. 451–464, 2008.

[24] S. C. de Greeff, A. L. M. Dekkers, P. Teunis, J. C. Rahamat-Langendoen, F. R. Mooi, and H. E. de Melker, "Seasonal patterns in time series of pertussis," *Epidemiology & Infection*, vol. 137, no. 10, pp. 1388–1395, 2009.

[25] R. Sparks, C. Carter, P. Graham et al., "Understanding sources of variation in syndromic surveillance for early warning of natural or intentional disease outbreaks," *IIE Transactions*, vol. 42, no. 9, pp. 613–631, 2010.

[26] W. A. Jensen, L. A. Jones-Farmer, C. W. Champ, and W. H. Woodall, "Effects of parameter estimation on control chart properties: a literature review," *Journal of Quality Technology*, vol. 38, no. 4, pp. 349–364, 2006.

[27] C. W. Champ, L. A. Jones-Farmer, and S. E. Rigdon, "Properties of the T^2 control chart when parameters are estimated," *Technometrics*, vol. 47, pp. 437–445, 2005.

[28] K. Hope, D. N. Durrheim, E. T. D'Espaignet, and C. Dalton, "Syndromic surveillance: is it a useful tool for local outbreak detection?" *Journal of Epidemiology and Community Health*, vol. 60, no. 5, pp. 374–375, 2006.

[29] Centers for Disease Control and Prevention, "Biosense 2," 2014, http://www.cdc.gov/biosense/.

[30] Department of Defense, *Essence (Electronic Surveillance System for the Early Notification of Community-Based Epidemics)*, 2014, http://health.mo.gov/data/essence/index.php.