Prospective Infectious Disease Outbreak Detection Using Markov Switching Models

Hsin-Min Lu, Daniel Zeng, *Senior Member*, *IEEE*, and Hsinchun Chen, *Fellow*, *IEEE*

Abstract—Accurate and timely detection of infectious disease outbreaks provides valuable information which can enable public health officials to respond to major public health threats in a timely fashion. However, disease outbreaks are often not directly observable. For surveillance systems used to detect outbreaks, noises caused by routine behavioral patterns and by special events can further complicate the detection task. Most existing detection methods combine a time series filtering procedure followed by a statistical surveillance method. The performance of this "two-step" detection method is hampered by the unrealistic assumption that the training data are outbreak-free. Moreover, existing approaches are sensitive to extreme values, which are common in real-world data sets. We considered the problem of identifying outbreak patterns in a syndrome count time series using Markov switching models. The disease outbreak states are modeled as hidden state variables which control the observed time series. A jump component is introduced to absorb sporadic extreme values that may otherwise weaken the ability to detect slow-moving disease outbreaks. Our approach outperformed several state-of-the-art detection methods in terms of detection sensitivity using both simulated and real-world data.

Index Terms—Markov switching models, syndromic surveillance, Gibbs sampling, outbreak detection.

✦

1 INTRODUCTION

DETECTING and controlling infectious disease outbreaks have long been a major concern in public health [1]. Recent efforts in building syndromic surveillance systems have increased the timeliness of the data collection process by incorporating novel data sources such as emergency department (ED) chief complaints (CCs) and over-the-counter (OTC) health product sales [2]. Research shows that these data sources contain valuable information that reflects current public health status [3], [4]. However, the noise caused by routine behavior patterns, seasonality, special events, and various other factors, is blended with the disease outbreak signals. As a result, disease outbreak detection using the time series from syndromic surveillance systems is a challenging task.

In a typical syndromic surveillance system [5], [6], [7], the data are classified and aggregated to generate univariate or multivariate time series at a daily frequency. An example of a univariate time series is the daily ED visits associated with a particular syndrome (for example, the respiratory syndrome). An example of a multivariate time series is the number of daily visits with a particular syndrome from multiple EDs. If geographic information such as the ZIP code is available, the multivariate time series in these examples would be the daily counts of patients with a particular syndrome from the ZIP code areas near an ED.

Most time series outbreak detection methods follow a two-step procedure [8], [9], [10]. In the first step, a baseline model describing the "normal pattern" is estimated using the training data that usually contain a historical time series without outbreaks. The baseline model then is used to predict future time series values. In the second step, statistical surveillance methods such as the Shewhart control chart [11], [12] or the Cumulated SUM (CUSUM) [13] method then take the prediction error (observed value minus predicted value) as the input and output alert scores. Higher alert scores are usually associated with a higher risk of having outbreaks. When the alert scores exceed a predefined threshold, the alarm is triggered.

Two main problems exist for current detection methods. First, the two-step procedure is based on the assumption that there are no outbreaks in the training data. When a real-world data set is used for training, the assumption is very hard to verify. Moreover, a full investigation of disease outbreaks during the data collection period is usually too expensive to conduct.

The validity of the detection results may be seriously impaired if it cannot be verified that the training data are outbreak-free. The estimated parameters of the baseline model may be biased by outbreak-related observations. Subsequent prediction and outbreak detection, as a result, may be negatively affected. The problem can seriously reduce the practical value of the outbreak detection method.

Second, existing time series detection methods also lack the ability to handle sporadic extreme values. Special events such as holidays and the media coverage of a particular disease may cause spikes that are not associated with disease outbreaks [14]. These extreme values usually last for a very short time (often just one or two days) and do not affect subsequent time series values. Anomalies related to real disease outbreaks, on the other hand, usually show a prolonged upward drift. The magnitude of disease-related

● *H.-M. Lu and H. Chen are with the Management Information Systems Department, University of Arizona, 1130 E. Helen Street, Tucson, AZ 85721-0108. E-mail: hmlu@email.arizona.edu, hchen@eller.arizona.edu.*
● *D. Zeng is with the Management Information Systems Department, University of Arizona, 1130 E. Helen Street, Tucson, AZ 85721-0108, and the Chinese Academy of Sciences, China. E-mail: zeng@eller.arizona.edu.*

drift is usually much smaller compared to the sporadic spikes caused by special events. Many outbreak detection algorithms take advantage of these characteristics and accumulate the errors so that small increases can be detected effectively [8], [9], [15]. The accumulation process, nevertheless, is susceptible to the presence of extreme values.

The deficiencies of current outbreak detection methods motivate our efforts to develop novel algorithms that can address these shortcomings. To deal with the problem of having outbreak-related observations in training data, a flexible statistical model must be used so that the model can adjust itself automatically when outbreak-related observations exist. In econometrics and time series literature, this is usually refereed to as the problem of modeling endogenous structural changes [16], [17].

A natural way of modeling structure changes in a time series is introducing additional hidden state variables which control the underlying time series dynamics. The Markov switching models originally proposed by Hamilton are one popular model of this kind [18]. This family of models includes a hidden state variable that may have a different value in each period. It takes values of either 0 or 1 that correspond to different conditional means, variances, and autocorrelations of the time series. The hidden state evolves following a first-order Markov process. That is, the current hidden state depends only on its historical values from the last period.

This hidden state method can be easily extended to handle extreme values. An additional hidden state can be included to model the presence of sporadic extreme values. With this additional hidden state, the model can distinguish between "normal" and "extreme" observations. That is, if a spike appears without signs that the sudden increase can be associated with drifts either before or after it, then the model can, based on the statistical evidence, assign the sudden increase as an extreme value instead of an outbreak. The negative effect of extreme values on outbreak detection can thus be reduced.

The main contribution of this paper is to present a prospective outbreak detection method that is robust to preexisting outbreaks and extreme values. Prospective outbreak detection, as opposed to retrospective detection in which the entire set of the observations is available to the detection algorithms, assumes that only observations made prior to the time of the detection are available to the detection algorithms. Retrospective detection is useful primarily for offline analysis of historical data, whereas prospective detection is intended for use in monitoring incoming public health data streams in an online fashion. We utilized the Markov switching model that includes three hidden state variables in each period. The first hidden state variable models the disease outbreak state, and the second hidden state variable models the presence of extreme value. If the extreme value exists, the third hidden state variable represents the size of the extreme value. We demonstrate that our approach outperforms several existing state-of-the-art outbreak detection algorithms using both simulated and real-world time series data.

This paper is organized as follows: Section 2 briefly introduces current outbreak detection methods and the Markov switching models. Section 3 presents our outbreak detection method. An evaluation study that uses both simulated and real-world data is summarized in Section 4. We conclude our paper in Section 5.

2 BACKGROUND

Current time series outbreak detection methods mostly follow a two-step procedure: a baseline time series estimation step followed by a statistical surveillance step [19], [8], [9]. We review these two major steps in this section.

Markov switching models, which belong to a broader class of statistical models that make use of hidden state variables, are also reviewed. We present the typical model settings and the estimation approaches.

2.1 Time Series Modeling

The first step in traditional outbreak detection methods is to develop a model that can describe the normal time series patterns. The most widely used model is the Autoregressive Integrated Moving Average (ARIMA) models of Box and Jenkins [20]. The model setting can be described by three parameters: (p, d, q). The parameter p refers to the length of historical time series values that can affect current observations. The second parameter d specifies how many different operations are required to make the time series stationary. The third parameter q specifies the length of historical error terms that can affect current observations. In a typical setting that does not involve seasonal fluctuation, the observed time series is usually assumed to be stationary, that is, $d = 0$. Specifically, an ARIMA$(p, 0, q)$ model can be written as:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} \cdots + a_p y_{t-p}$$
$$+ \epsilon_t + b_1 \epsilon_{t-1} \cdots + b_q \epsilon_{t-q},$$

where y_t is the observed time series and ϵ_t is the error term. To ensure that the model "learns" the normal time series pattern, the data used for model estimation should be outbreak-free. Given p and q, the parameter values $(a_0, a_1, \ldots, b_q)$ can be estimated using likelihood maximization [21]. However, different model settings that correspond to different values of p and q may affect prediction accuracy. The values of p and q are usually determined by model selection criteria that take both goodness of fit and model complexity into consideration. Commonly used model selection criteria include Akaike information criterion (AIC) [22], [23] and Bayesian information criterion (BIC) [24]. Note that the model selection criteria are closely related to the "cross-validation" evaluation approach [25] commonly used by the machine learning community [26]. In fact, cross-validation is asymptotically equivalent to AIC [27].

Other modeling techniques such as the generalized linear model using Poisson distribution [28], expectation-variance model [29], and the Wavelet model [30] have been evaluated in previous studies.

For the purpose of detecting outbreaks, there are two issues warranting further discussion: the modeling of the day-of-week and seasonal effects.

2.1.1 Day-of-Week Effect

The syndromic surveillance time series usually exhibits strong day-of-week effects. For example, there are usually more ED visits during the weekends than during the weekdays [9]. The variation among different day-of-weeks is usually assumed to be fixed. As an illustrative example, an ARIMA$(1,0,0)$ model with a fixed day-of-week effect can be written as

$$y_t = w_1 d_{t,1} + w_2 d_{t,2} \cdots + w_6 d_{t,6} + a_0 + a_1 y_{t-1} + \epsilon_t,$$

where $d_{t,i} \in \{0,1\}$, $i = 1,2,3,4,5,6$ are dummy variables indicating a particular day-of-week. For example, $d_{t,1} = 1$ if day t is a Monday and 0 otherwise. Note that we need only six dummy variables for seven day-of-weeks because of the existence of the constant term a_0.

2.1.2 Seasonal Effect

Similar to the day-of-week effect that refers to a weekly cyclic pattern, the seasonal effect refers to a yearly cyclic pattern. Tri-geometric functions are commonly used to model deterministic seasonal fluctuation. This technique is usually referred to as the Serfling model [31], [32], which can be written as:

$$y_t = a_0 + b_1 \, cos\left(\frac{2\pi t}{365.25}\right) + b_2 \, sin\left(\frac{2\pi t}{365.25}\right)$$
$$w_1 d_{t,1} + w_2 d_{t,2} \cdots + w_6 d_{t,6} + \epsilon_t.$$

Note that both day-of-week and seasonality are included in the model. The model can be refined by including more tri-geometric functions that correspond to semiannual and even quarterly cyclic patterns. However, it has the obvious problem of assuming the same seasonal peaks and troughs across the whole monitoring period [32]. Our preliminary experiments show that the Serfling model fits the observed syndromic time series poorly especially when the seasonality is strong. The Serfling model assumes a particular shape of the time series that may not be empirically valid.

Other modeling techniques allow more flexible seasonal fluctuation across years. One possibility is to use the Holt-Winters exponential smoothing to model seasonality [33], [34]. An empirical study showed that, in the context of syndromic surveillance, Holt-Winter exponential smoothing outperformed the Serfling model in terms of prediction accuracy [35].

The concept of the seasonal random walk [36] can be applied to model the seasonal effect. The basic idea is that the same day-of-year should have the same expected value. Reis and his colleagues estimated the expected value using the trimmed-mean of historical time series value with the same day-of-year in an 8-year window [8], [9]. The seasonal effect can then be filtered out by subtracting the observed value from the day-of-year expectation.

2.2 Statistical Surveillance Methods

For outbreak detection purposes, the prediction errors from the time series modeling step are further processed using statistical surveillance methods. Various statistical surveillance methods such as the Shewhart control Chart [11], CUSUM [13], Exponential Weighted Moving Average (EWMA) [12], Shiryaew-Roberts method [37], [38], and the likelihood ratio methods [39] can be applied for disease outbreak detection. However, most syndromic surveillance studies use the Shewhart control chart, CUSUM, EWMA, and their variations. Our review focused mainly on these three methods. For more details, we refer the reader to [40].

The Shewhart control chart [11] checks the t-value of the prediction errors period by period. It performs the best if large, isolated outbreaks are involved. However, since disease outbreaks often exhibit only small deviations in their early stages, the Shewhart control chart may not be the best choice for our purposes.

The CUSUM method minimizes the maximum value of the conditional expected delay "when the outcome before outbreak is the worst possible" [41]. It uses a recursive formula to accumulate the prediction errors:

$$C_t^+ = max[0, e_t - K + C_{t-1}^+],$$

where e_t is the prediction error from the time series model and K is a predefined constant that is commonly referred to as the allowance. The alarm is triggered if C_t^+ exceeds a predefined threshold.

The EWMA method can be seen as a linear approximation of the likelihood ratio method [39], [42]. The alert score is computed by accumulating forecasting errors with exponentially decaying weights. Similar to the CUSUM method, higher outbreak scores are usually associated with a higher risk of having an outbreak. The threshold can be determined from theoretical analysis or empirical studies [43], [44].

Some syndromic surveillance studies use a moving average scheme to accumulate forecasting errors [8], [9]. Their studies have showed that a linear-increasing weighting schemes performed the best in terms of outbreak detection ability.

2.3 Performance Measures

The most commonly used performance measure in statistical surveillance literature is the Average Run Length (ARL). ARL^0 denotes the expected run length until the first false alarm, and ARL^1 denotes the expected run length until an alarm when the process is out of control at the start of the surveillance [40], [45], [44].

These measures, nevertheless, are less intuitive under the context of disease outbreak detection. Most disease outbreak detection studies use per-day sensitivity and false alarm rate (FAR) [28], [29], [8]. Sensitivity is the probability of having alarms on outbreak days. False alarm rate is the probability of having alarms on nonoutbreak days.

2.4 Extreme Values in Syndromic Surveillance Time Series

Current surveillance methods are very sensitive to extreme values. The main reason is because the statistical surveillance methods accumulate the forecasting errors and there are no simple methods that can be used to filter out the extreme values. Burkom [46] proposed using a "reset" rule to bring down the alert scores when extreme values are known to be causing the elevated scores. However, it is not clear how to establish effective reset rules.

Common reasons behind the extreme values include holidays, media coverage, and special events [14]. However, existing studies have not offered help for handling the
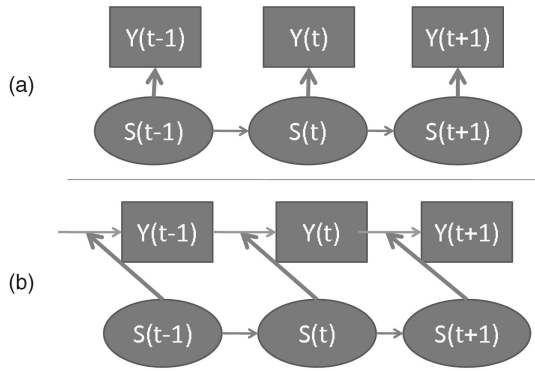
Fig. 1. (a) Hidden Markov models and (b) Markov Switching models. The rectangles are observable random variables and the circles are hidden state variables. Arrows indicate the dependencies among variables.

negative effects caused by extreme values. Previous studies have used holiday dummies to absorb the holiday effects [29]. This technique, nevertheless, imposes an unrealistic assumption that all holidays have the same effect on the time series.

2.5 The Markov Switching Model

The Markov switching model belongs to the family of state-space models. A state-space model is a statistical model with hidden state variables controlling observable random variables. There are two types of equations in this model: the measurement equations and the transition equations [47]. The measurement equation defines how hidden states affect the observable random variables. The transition equation, on the other hand, defines how the state variables evolve over time.

When the state variable is discrete, the state-space model is usually called the hidden Markov model [48], [49] or the Markov switching model [18] depending on the choice of the measurement equation. The measurement equation in the hidden Markov model is usually formulated so that the observable random variables at period t only depend on the hidden state variables at the same period.

The Markov switching model addresses the weakness of the hidden Markov model by including lagged observations. The observable random variables in the Markov switching model depend on their historical values as well as the hidden state variables. This setting makes the Markov switching model more suitable for time-series-related problems. Fig. 1 illustrates the dependency difference between the Markov switching model and the hidden Markov model.

Strat and Carrat [50] applied the state-space model for disease outbreak detection. They used a two-state hidden Markov model on a weekly influenza-like illness (ILI) incidence and showed that the hidden Markov model clearly differentiated between epidemic and nonepidemic rates. However, as they pointed out in the conclusion, "the validity of the hypothesis that ILI incidence rates are independent conditional on the state, is questionable." They also pointed out that autoregressive terms should be included for better performance. We are unaware of prior studies on applying Markov switching models for outbreak detection.

Most applications of the Markov switching models fall in the field of economics and finance. Notable examples are identifying macroeconomics business cycle [18] and modeling changing interest rates regimes [51]. A simple Markov switching model can be written as

$$y_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1} + e_t, \qquad (1)$$

$$p(s_t = j | s_{t-1} = i) = p_{ij}, \qquad (2)$$

$$s_t \in \{0, 1\}, \qquad (3)$$

$$e_t \sim N(0, \sigma^2). \qquad (4)$$

Equation (1) defines how the hidden state variable s_t controls the dynamics of the observable random variable y_t. At an nonoutbreak period ($s_t = 0$), y_t is determined by a drift term $a_{0,0}$ and the autoregressive parameter $a_{1,0}$. If an outbreak occurs ($s_t = 1$), the drift term increases to $a_{0,0} + a_{0,1}$ and the autoregressive parameter increases to $a_{1,0} + a_{1,1}$ (assuming $a_{0,1} \geq 0$ and $a_{1,1} \geq 0$). Equation (2) indicates that the hidden states evolve following a Markov process with transition probability p_{ij}.

Note that if we have a time series of T period, there are four parameters and T hidden state variables in (1), together with two transition probability in (2) and a variance for error terms in (4). We have more unknowns than the number of periods, which complicates the estimation process. We briefly discuss the model estimation issues below.

2.5.1 Model Estimation for the Markov Switching Model

Model estimation for the Markov switching model is much more complicated than that of the standard time series models such as the ARIMA models. The technical difficulty arises from the presence of unknown hidden states. In a simplified case involving only one hidden outbreak state variable with two possible states and a total of T periods, a direct evaluation of the likelihood function involves a summation of all possible trajectories of hidden states. The time complexity is $O(2^T)$, which is intractable in practice. More sophisticated algorithms, which compute the posterior distribution of the hidden states using a forward-filtering-backward-smoothing (FFBS) procedure [47], [52], take only $O(2^3T)$ steps. The computation of the posterior distribution of the hidden states is required by many estimation methods such as the expectation-maximization (EM) algorithm [53], [54], [55], Gibbs sampling, and Markov Chain Monte Carlo (MCMC) [56], [57], [58]. Note that to deliver the final optimal parameter estimation, these algorithms need to execute repeatedly until certain convergence criteria are met.

The EM algorithm finds the maximum of the likelihood function by iterating between calculating the expected value of state variables given current parameters and calculating the maximum of log likelihood given the expected state variables. It was applied to estimate the hidden Markov model in a previous outbreak detection study [50]. Compared to other numerical optimization methods, the EM algorithm is more robust and usually converges if a maximum exists. However, it is possible that

the algorithm converges to a local maximum instead. In practice, the EM algorithm is run with multiple initial values.

A serious drawback of the EM algorithm is the label switching problem [52]. The Markov switching model (and the hidden Markov model) is invariant under arbitrary permutations of the state labels. As a result, we cannot be sure whether $s_t = 0$ is representing an outbreak or non-outbreak state before the estimation procedure is completed. The label switching problem is especially an issue when the Markov switching model is part of a larger automatic disease outbreak detection system.

Gibbs sampling [57], [58], [59] is an alternative estimation method that can avoid the label switching problem. The Gibbs sampling iterates to draw random variables from conditional posterior distributions of parameters and state variables to simulate the full posterior distribution of parameters and state variables. Specifically, let $\Theta = \{\theta_1, \ldots, \theta_k\}$ denote the unknown parameters (and state variables). By the Bayes Theorem, the posterior distribution $p(\Theta|Y)$ is proportional to the likelihood of $p(Y|\Theta)$ multiplying the prior of parameters $p(\Theta)$. The label switching problem can be avoided by imposing proper constraints on $p(\Theta)$. Gibbs sampling estimates parameters using a simulation-based method. The following steps can be used to simulate Θ from its posterior distribution. First, select initial values $\Theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_k^{(0)}\}$. For $i = 1, 2, \ldots I$, iterate through the following steps:

1. Draw $\theta_1^{(i)}$ from

$$p\big(\theta_1|Y, \theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)}\big).$$

2. Draw $\theta_2^{(i)}$ from

$$p\big(\theta_2|Y, \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}\big).$$

...

3. Draw $\theta_k(i)$ from

$$p\big(\theta_k|Y, \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{k-1}^{(i)}\big).$$

4. Record $\Theta^{(i)} \equiv \{\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)}\}$.

It has been shown that $\{\Theta^{(i)}\}$ converges to $p(\Theta|Y)$ [60], [61]. As a result, the posterior mean of θ_j can be estimated by the average of $\{\theta_j^{(i)}\}$, excluding certain "burn-in" iterations to minimize the effect of the initial value. The confidence intervals of the estimated parameters can also be calculated directly from $\{\theta_j^{(i)}\}$.

3 OUTBREAK DETECTION USING THE MARKOV SWITCHING WITH JUMPS (MSJ) MODEL

We developed our disease outbreak detection algorithm based on the Markov switching models [18]. Two hidden disease outbreak states (0 or 1; nonoutbreak or outbreak) were assumed. To handle the sporadic extreme values, we included a jump component to filter their negative effects on outbreak detection. Seasonality was handled based on the concept of seasonal random walk.

Our proposed MSJ model is described below:

$$y_t = g(Y^{t-1}) + z_t, \tag{5}$$

$$z_t = \xi_t J_t + x_t, \tag{6}$$

$$x_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)x_{t-1} + \sum_{i=1}^{6} w_i d_{t,i} + \sum_{i=1}^{K} b_i v_{t,i} + e_t, \tag{7}$$

$$s_t \in \{0, 1\}, \tag{8}$$

$$J_t \in \{0, 1\}, \tag{9}$$

$$p(s_t = j|s_{t-1} = i) = p_{ij}, \tag{10}$$

$$e_t \sim N(0, \sigma^2), \tag{11}$$

$$\xi_t \sim N\big(0, \sigma_a^2\big), \tag{12}$$

$$g(Y_{t-1}) = med\{\bar{y}_{t-m}, \bar{y}_{t-2m}, \bar{y}_{t-3m}\}, \tag{13}$$

$$\bar{y}_{t-im} = \frac{(y_{t-im-3} + y_{t-im-2} \cdots + y_{t-im+3})}{7}, \tag{14}$$

where $Y^{t-1} = (y_1, y_2, \ldots, y_{t-1})$ and $m = 365$. The hidden state variable $s_t = 1$ indicates period t is an outbreak period, 0 otherwise. In the subsequent discussion, the time period in subscript indicates the value at that period; the time period in superscript indicates a vector of values up to that period.

Equation (5) filters out the seasonal fluctuation by subtracting the day-of-year expectation from observed time series values. The day-of-year expectation is estimated using the historical values within a day-of-year window in the past three years ((13) and (14)). The next equation, (6), further decomposes the residual (z_t) into normal variation (x_t) and a possible jump component. If a jump exists ($J_t = 1$), then ξ_t is the size of the jump. Equation (7) articulates the dynamic behavior during outbreak and nonoutbreak periods. The hidden state variable s_t controls the constant term and an autoregressive coefficient. The variables $d_{t,i}$ are day-of-week dummies. The exogenous variables $v_{t,i}$ are optional controlling factors. Environmental variables such as pollen level and temperature are two possibilities. If necessary, more lagged dependent variables can also be included. For example, we can set $v_{t,1} = x_{t-2}$, $v_{t,2} = x_{t-3}, \ldots, v_{t,6} = x_{t-7}$. As defined in (10), the transition of s_t follows a first-order Markov process.

Compared to conducting outbreak detection using a baseline time series model combined with a statistical surveillance method, our approach provides the following advantages: First, the alert scores ($p(s_t = 1|Y_t)$) of our approach have a clear and intuitive interpretation. Most existing outbreak detecting methods output alert scores that do not have clear meanings. The only way to make sense of the alert scores is to compare the scores with an established threshold. The alert score of our detection algorithm, without referring to any thresholds, can be interpreted as the outbreak probability given available information.

Second, our algorithm provides an estimated outbreak size in addition to outbreak probability. In traditional outbreak detection methods, it is not easy to estimate the outbreak size directly from the alert statistics or estimated parameters. Our method allows the model to recognize different temporal dynamics in different periods. The outbreak size can be calculated directly from the estimated parameters. The information could be valuable for the planning of public health intervention.

Third, the jump component gives our algorithm the ability to separate sporadic extreme values from slow-moving disease outbreaks. The additional information provides flexibility that is valuable for different surveillance needs.

3.1 Changing Dynamics and Outbreak Size

The hidden variable s_t plays an important role in determining the dynamics of x_t. Consider a simplified setting with no day-of-week effect $(w_i = 0)$ nor exogenous variables $(b_i = 0)$. If we have $s_t = 0$ for all time except $t = t_1$, then the observed value increases by $\Delta_{t_1} \equiv a_{0,1} + a_{1,1} y_{t_1-1}$ at t_1, ignoring the effect of the noise (e_t). Note that the autoregressive coefficient $a_{1,1}$ also plays a role in determining the magnitude of the increase at time t_1. After this time point, the effect of Δ_{t_1} decreases exponentially. The scenario is similar to dropping a group of infected persons in a large community at period t_1 and seeing the disease starting to spread. However, since infected persons recover from the disease quickly, the disease dies out quickly as well.

If $s_t = 1$ for $t = t_1, t_1 + 1, \ldots$, the effect of increased constant term and autoregressive coefficients accumulates during the outbreak periods until it reaches the new stable level. The new long-term mean can be found by writing x_t as a function of $a_{i,j}$ and e_t only. A simple computation gives $E[x_t|s_t = 1] \equiv \bar{m}_2 = (a_{0,0} + a_{0,1})/(1 - a_{1,0} - a_{1,1})$. Similarly, the long-term mean of nonoutbreak periods is $E[x_t|s_t = 0] \equiv \bar{m}_1 = a_{0,0}/(1 - a_{1,0})$. The outbreak size is the difference between $\bar{m}_2$ and $\bar{m}_1$.

3.2 Model Estimation

Gibbs sampling is used for model estimation. We need to estimate the following sets of coefficients and hidden states: time series coefficients $A = (a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1})$, day-of-week coefficients $W = (w_1, w_2, \ldots, w_6)$, exogenous variable coefficients $B = (b_1, b_2, \ldots, b_k)$, variance of the error term (σ^2), transition probability $P = (p_{00}, p_{11})$, hidden outbreak state $S^T = (s_1, s_2, \ldots, s_T)$, hidden jump state $J^T = (J_1, J_2, \ldots, J_T)$, hidden jump size $\Xi^T = (\xi_1, \xi_2, \ldots, \xi_T)$, and variance of jumps (σ_a^2).

For a model with T periods and K exogenous variables, there are $3T + K + 14$ parameters to be estimated. Note that most of these parameters are subjected to certain constraints. For example, the outbreak states are assumed to follow a first-order Markov process. Given the transition probabilities and the observed time series, some parameter combinations are almost impossible or have a very low probability. The goal of model estimation is to search the parameter space and locate the area that is corresponding to a high probability given observed time series.

To facilitate the simulation of random variables from the posterior distributions, conjugate priors are used for all parameters. As discussed in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.115, all

TABLE 1
Conditional Posterior Distributions

$(A, W, B)\|\bullet \sim$ Multivariate Normal
$\sigma^2\|\bullet \sim$ Inverse Gamma
$\xi_t\|\bullet \sim$ Normal
$J_t\|\bullet \sim$ Binomial
$s_t\|\bullet \sim$ Binomial
$\sigma_a^2\|\bullet \sim$ Inverse Gamma
$p_{ii}\|\bullet \sim$ Beta

conditional posteriors follow well-known statistical distributions and are summarized in Table 1. The dot ($\bullet$) in Table 1 indicates the conditioning on other parameters and hidden states. To increase the efficiency of sampling s_t, the FFBS procedure is used.

It should be noted that to avoid the label switching problem, we constraint the parameter sampling results so that $\bar{m}_1 < \bar{m}_2$ is satisfied. If the constraint is violated (A, W, B) are redrawn until the constraint is satisfied.

3.3 Prospective Outbreak Detection

Given an up-to-date time series, prospective outbreak detection answers the question "What is the probability of having a disease outbreak today?" Letting t denote the current time period, we want to estimate $p(s_t = 1|Y^t)$, where s_t is the hidden outbreak state and Y^t is the vector contains all time series values up to time t. When a new time series value arrives in the next period, the system needs to rerun the model and provide the estimation of $p(s_{t+1}|Y^{t+1})$.

Our preliminary experiments found that direct implementation of the estimation algorithm provides little valuable outbreak information because the algorithm became too sensitive to small changes. The algorithm tried to scrutinize all small changes and tended to overreact to those changes. To overcome this difficulty, we developed a regulation technique to desensitize the algorithm so that small, unimportant changes would be ignored.

3.3.1 Desensitization for Prospective Outbreak Detection

The desensitization technique is an extension of the solution for the label switching problem. To make the algorithm ignore small, unimportant changes, we rejected the parameter sampling results that indicated small changes. Specifically, we chose g as the minimal outbreak size that we wanted to detect. We let $a_{0,0}^{(c)}, a_{0,1}^{(c)}, a_{1,0}^{(c)}, a_{1,1}^{(c)}$ be the sampling result of the cth iteration. We rejected the sampling result if $\bar{m}_1^{(c)} \geq \bar{m}_2^{(c)} - g$. The coefficient g is set to 5 percent of the time series mean during the training period. Also, the autoregressive coefficient needs to have a value between -1 and 1 to ensure that the time series is stationary. The desensitization procedure is summarized in Algorithm 1.

Algorithm 1. Desensitization Procedure
  repeat
    Draw $(A^{(c)}, B^{(c)}, W^{(c)})$ from $(A, B, W)| \bullet$.
    $\bar{m}_1^{(c)} \leftarrow a_{0,0}^{(c)}/(1 - a_{1,0}^{(c)})$.
    $\bar{m}_2^{(c)} \leftarrow (a_{0,0}^{(c)} + a_{0,1}^{(c)})/(1 - a_{1,0}^{(c)} - a_{1,1}^{(c)})$.
  until $\bar{m}_1^{(c)} < \bar{m}_2^{(c)} - g$ and $|a_{1,0}| < 1$ and $|a_{1,0} + a_{1,1}| < 1$.
  return $(A^{(c)}, B^{(c)}, W^{(c)})$.

TABLE 2
Prior Distributions

Parameter	Distribution	Parameter
$\{a_{0,0}, a_{0,1},$ $a_{1,0}, a_{1,1}\}$	Multivariate Normal(M, V)	$M = \{0, 0, 0.15, 0.6\}$ $\{v_{ii}\}_{i=1}^4 = \{400, 400, 3, 3\}$
σ^2	Inverse Gamma	$\alpha = 3, \beta = $ est. variance $\times (\alpha - 1)$
σ_a^2	Inverse Gamma	$\alpha = 3, \beta = $ est. variance $\times 5(\alpha - 1)$
$\{w_1, ..., w_6\}$	Multivariate Normal(M, V)	M is est. from the training data $\{v_{ii}\}_{i=1}^6 = \{v_a, v_a, v_a, v_a, v_a, v_a\}$; $v_a = \max(100, 5\max(M))$
p_{11}	Beta	a=2, b=0.2
p_{22}	Beta	a=2, b=0.1

The off-diagnose elements of V is set to zero

Note that the presence of outbreaks in the training period may bias the estimated mean and also lead to a larger coefficient g. As a result, the model becomes less sensitive to outbreaks. We argue that, with a moderate length of training period (say, one year), the proposed method of choosing g is robust against the presence of possible outbreaks. The main reason is that, in most cases, actual outbreak periods are much shorter than the training period, and taking average during the training period can dilute the effect of outbreak observations to a large extent. As an example, assume that the mean of daily counts is 100 and the daily count increases to 200 (a 100 percent increase) during an one-month outbreak. The chosen g will increase for less than 10 percent, from $5(100 \times 0.05)$ to $5.4((100 \times 30/365 + 200 \times 335/365) \times 0.05)$. Our computational experience suggests that small variations in g do not have a significant effect on the detection outcome. On the other hand, in cases where the training period is relatively short (say, 3 months), we caution that the negative impact of setting unrealistic values for the counts and outbreak size can be noticeable from the point of view of detection power. In these cases, domain experts' input on how to set g will be highly valuable.

3.3.2 Prior Distributions

While some parameters of the prior distributions are quite robust to various circumstances, others need to be customized case by case. We applied a simple AR(1) model with day-of-week effect on the training data with seasonality removed. The estimated variance of the error term is used to set up the parameters for the prior of σ^2 and σ_a^2. The estimated day-of-week effects are used to set up the prior of w_i. The prior distributions used in this study are summarized in Table 2.

3.3.3 Summary of the Estimation Procedure

Given a time series covering period 1 to t_1, our goal is to estimate the outbreak probability of period t_1, together with other relevant parameters and hidden state variables. Using Gibbs sampling for estimation, we need to choose the total number of iteration B and the "burn-in" iteration b. The sampling results between iteration $b+1$ and B are then used to compute the outbreak probability (alert score) and the estimates of other parameters. The pseudocode that summarizes the procedure can be found in Algorithm 2. We implemented our approach on R, an open-source statistical software (http://www.r-project.org/).
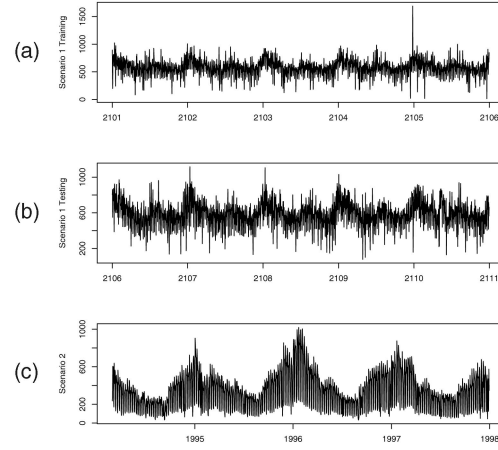

Fig. 2. (a) The training data of Scenario 1. (b) One of the testing data of Scenario 1. (c) The original real-world time series used in Scenario 2.

Algorithm 2. Prospective Outbreak Detection Using the Markov Switching with Jumps (MSJ) Model

for $c = 1$ to B **do**
$(A^{(c)}, B^{(c)}, W^{(c)}) \leftarrow$ Desensitization().
Draw $\sigma^{2(c)}$ from $\sigma^2 | \bullet$.
Draw $s_{t_1}^{(c)}, s_{t_1-1}^{(c)}, \ldots, s_1^{(c)}$ using FFBS.

Draw $J_t^{(c)}$ from $J_t | \bullet$ for $t = 1, 2, \ldots, t_1$.

Draw $\xi_t^{(c)}$ from $\xi_t | \bullet$ for $t = 1, 2, \ldots, t_1$.
Draw $\sigma_a^{2(c)}$ from $\sigma_a | \bullet$.
end for
$\hat{p}(s_{t_1} = 1 | Y^t) \leftarrow \sum_{c=b+1}^{B} s_{t_1}^{(c)} / (B - b + 1)$.

4 EVALUATION STUDY

We developed two disease outbreak scenarios to evaluate our approach. Scenario 1 is aggregated OTC antidiarrheal and antinauseant sales simulated based on a real-world data set developed by the International Society for Disease Surveillance (ISDS) [62]. The outbreaks in this scenario were simulated based on "a large waterborne outbreak of Cryptosporidium [which] occurred in the Battleford area of Saskatchewan during the spring of 2001. Due to the prolonged, less severe nature of Cryptosporidium, many infected residents self-medicated, evidenced by an increase of OTC antidiarrheal and antinauseant product sales during the outbreak." (cf. https://wiki.cirg.washington.edu/pub/bin/view/Isds/TechnicalContest).

This data set contains 5 years of training data and 30 5-year time series data sets with outbreaks for model testing. The starting date of training data is marked "1/1/2101." The training data and one of the 30 testing time series are plotted in Figs. 2a and 2b. Note that the plotted testing data set contains an outbreak starting from "4/15/2110" that lasts for 54 days.

Scenario 2 used a real-world clinic visit time series and simulated outbreaks following the standard approach widely used in the syndromic surveillance literature [35], [8], [9]. We imposed simulated anthrax outbreaks on the clinic visit time series collected from a metropolitan area. The clinic visit is classified into syndromes using ICD-9 code

TABLE 3
Parameters for Anthrax Outbreak Simulation:
Disease Progression

Parameter	Value	Source
Incubation duration, median	11 days	[63]
Incubation duration, dispersion	2.04 days	[63]
Prodromal duration, median	2.50 days	[63]
Prodromal duration, dispersion	1.44 days	[63]

TABLE 4
Parameters for Anthrax Outbreak Simulation:
Healthcare Seeking at Prodromal State

Parameter	Value	Source
Prob. of seeking care	0.4	[15]
Prob. of respiratory syndrome	0.7	[64]
Prob. of gastrointestinal syndrome	0.2	[64]
Prob. of fever syndrome	0.1	[64]

according to the definitions from CDC (cf. http://www.bt.cdc.gov/surveillance/syndromedef/word/syndromedefinitions.doc).

The data set covers the period from 2/28/1994 to 12/30/1997 with a total of 1,402 days. Observations before 12/31/1995 were reserved for model training. Outbreak periods were randomly chosen between 1/1/1996 and 12/30/1997. Since the respiratory syndrome is the most common syndrome for early infection of inhalational anthrax, we focused on detection disease outbreak using aggregated clinic visits with the respiratory syndrome in this study. Fig. 2c plots the respiratory syndrome count time series used in this study.

We chose $B = 300$ and $b = 100$. Preliminary experiments on simulated data showed that this setting was appropriate for estimating the hidden states and parameters.

We discuss our outbreak simulation method for Scenario 2 in more detail and benchmark surveillance methods for both Scenarios in sequence. Section 4.4 presents experimental results.

4.1 Simulating Disease Outbreaks for Scenario 2

There are two major components in simulating the disease outbreak caused by inhalational anthrax [15]. The first component is the disease progression of infected persons [63]. The second component is the healthcare seeking behavior of infected persons. Since we aimed to focus on temporal disease outbreak detection, the spatial dispersion of anthrax spores and infection in different areas [63] were not considered in the simulation.

At the beginning of the simulation, we assumed that, in total, there are S infected persons. For each infected person, the disease progresses through three states: incubation, prodromal, and fulminant. The length of each state follows a log-normal distribution. Disease symptoms start to appear in the prodromal state. An infected person may have respiratory, gastrointestinal, or fever syndromes in the prodromal state. When the infected person is in the fulminant state, the person may exhibit shock syndrome or neurological syndrome. Since syndromes in the fulminant state are not the focus in this study, this state is not simulated. Parameters used in the simulation are summarized in Tables 3 and 4.

We set S to 15,000. This setting corresponds to an average peak of 566 patients. The outbreak period begins when anthrax spores are released and ends when more than 90 percent of infected persons with the respiratory syndrome have realized. Since there are usually a small number of patients with a long incubation period, the outbreak period may be artificially long just because few persons have late onset of the syndrome. The 90 percent

cutoff ensures that the outbreak period covers the most intense period of anthrax outbreaks.

The outbreak signals were imposed on the real-world time series with a starting time chosen randomly between 1/1/1996 and 12/30/1997. We generated 50 synthetic data sets for evaluation, each containing one simulated outbreak.

4.2 Benchmark Temporal Detection Methods

We chose the Serfling model with the CUSUM method as our first benchmark detection method. The Serfling model is one of the most popular time series models that incorporate seasonal fluctuations [31], [32]. The model can be written as follows:

$$y_t = a_0 + \sum_{i=1}^{6} d_{t,i} w_i + b_1 cos(2\pi t/365.25)$$
$$+ b_2 sin(2\pi t/365.25) + e_t,$$
$$e_t \sim N(0, \sigma^2),$$

where $d_{t,i}$ is day-of-week dummy. Note that we only need six day-of-week dummies.

Specifically, let $t = 1$ denote the starting day of the testing period. When the detection system started, one-step-ahead prediction (for $t = 1$) was made with parameters trained using counts up to $t = 0$. The standardized prediction error was fed into the CUSUM method. Then, the time advanced by one day. The count corresponding to $t = 1$ then was made available to the model. The one-step-ahead prediction for $t = 2$ was made with parameters trained using all the counts available up to $t = 1$. The process was repeated until the end of the testing period. This benchmark detection method is referred to as the S+CUSUM method in the subsequent discussions.

The second benchmark method, the trimmed-mean seasonal ARMA model, was implemented following [9], [8]. Observations made in the training period were used to estimate the overall mean, the mean for day-of-week, and the trimmed-mean for day-of-year. For observations in the testing periods, the overall mean, the mean for day-of-week, and the trimmed-mean for day-of-year were subtracted from the raw count. The demeaned counts were then fed into an ARMA(p,q) to filter out high-frequency dependency. We chose $p = 1$ and $q = 0$ for Scenario 1 and $p = 7$ and $q = 0$ for Scenario 2 according to Akaike Information Criteria (AIC). One-step-ahead predictions were calculated and prediction errors from the ARMA model were then weighted according to a linear-increasing pattern to compute the alert score. New observations were included for parameter estimation when available. This method is referred to as T+MA (Moving Average) and our approach is referred to as MSJ in subsequent discussion.
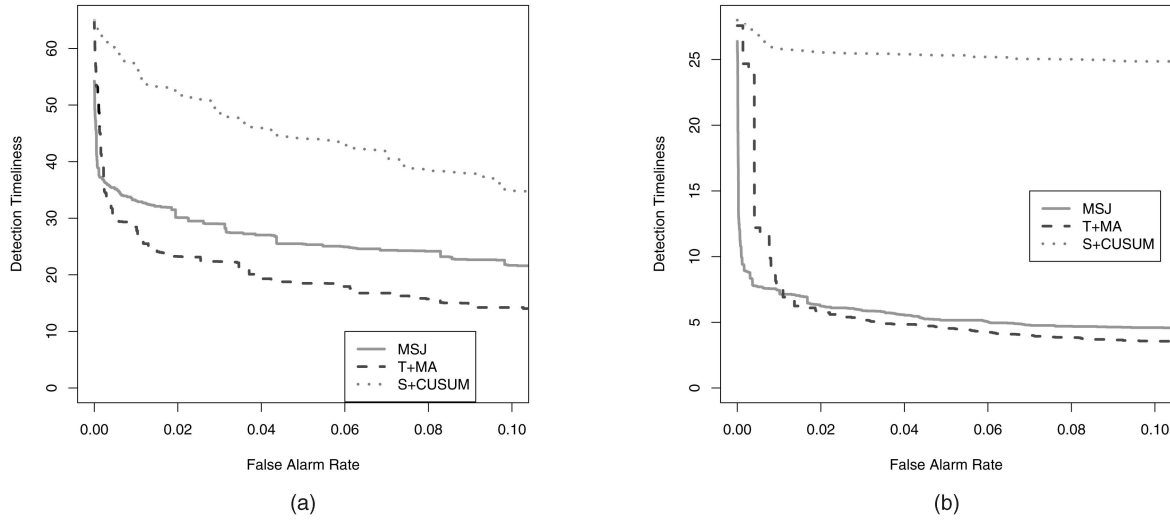
Fig. 3. Performance Comparison: Timeliness. (a) Scenario 1 detection timeliness at different false alarm rates. (b) Scenario 2 detection timeliness at different false alarm rates.

Note that we have adopted an "expanding window" or "rolling horizon" approach for the baseline models as well as our MSJ model. For a given day, all historical data were used for model training and the most up-to-date parameters were used for prediction and detection. The original training period (as mentioned in the beginning of Section 4) was mainly used for seasonality filtering, model selection (for T+MA) and prior distributions setting (for MSJ). The advantage of this expanding window approach is the efficiency of using information available to the underlying model.

4.3 Evaluation Metrics

We use two common syndromic surveillance evaluation metrics in this study [32], [65], [66]. The first metric is detection timeliness [62]. It measures the delay from the onset of the disease outbreak to the first detection of the disease outbreak at a given level of false alarm rate. If an outbreak is not detected across the whole outbreak period, the delay time is counted as the maximum outbreak length in all testing runs (65 days for Scenario 1, and 28 days for Scenario 2).

The FAR is defined as the probability of having an alarm for nonoutbreak days [28], [29], [8], [9]. For example, an FAR of 0.1 means that, on average, there are about $365 \times 0.1 = 36$ days with false alarms in a year with no outbreaks.

The second metric is per-day detection sensitivity [9]. This metric measures the probability of detecting an outbreak on an outbreak day. Given an alert threshold h, let o_h be the number of outbreak days that have alert scores exceeding h and Q be the total outbreak days in the testing data set, the detection sensitivity is o_h/Q.

4.4 Results

Figs. 3a and 3b plot the detection timeliness of Scenarios 1 and 2 at different false alarm rates. The solid line corresponds to the delay of our approach. Dashed and dotted lines indicate the T+MA and S+CUSUM methods. We consider the false alarm rate only when it is less than or equal to 0.1. A detection system with a false alarm rate higher than 0.1 is usually considered impractical because of the high cost

associated with confirming the false alarms. As clearly observed in the figures, our approach started with the lowest detection delay compared to other benchmark methods. As the FAR increased, the delay decreased for all methods. The detection delay of our approach remained the lowest for a range of FAR and then the T+MA method became the lowest.

The S+CUSUM method had the worst detection speed. The detection delay is at the maximum outbreak length (65 days for Scenario 1, and 28 days for Scenario 2) when the FAR is 0. It indicated that the S+CUSUM method detected no outbreaks when no false alarms were allowed. The improvement of detection delay was the smallest among all methods when FAR increased.

To further analyze the detection delay, we have tested the null hypothesis that our approach and the benchmark methods have the same detection delay. The last two columns of Table 5 report the differences in detection delay and the p-values (in the parenthesis) for two-tailed hypothesis testing. The p-value was computed using a bootstrap procedure (see, e.g., [67]). Our approach has lower detection delay than the S+CUSUM method. The difference is, in most cases, significant at conventional confidence levels for Scenario 2 but not Scenario 1. On the other hand, our method has higher detection delay compared to the T+MA method except for the case of $FAR = 0$. The difference, nevertheless, is not significant. The results indicated that the detection timeliness of our approach is better than that of the S+CUSUM method but is at the same level as that of the T+MA method.

Figs. 4a and 4b plot the detection sensitivity of all detection methods under the FAR we considered. Our approach had the highest detection sensitivity compared to the benchmark methods. The T+MA came in second, followed by the S+CUSUM method. The performance gaps remained consistent at FAR greater than 0.0125. In fact, the sensitivity of our approach was on average 0.15 higher than the T+MA method in Scenario 1 and 0.09 higher in Scenario 2. The performance gap was even larger (0.28 and 0.26) compared with the S+CUSUM method. In some FARs, the gaps represented a relative difference of more than 100 percent.

TABLE 5
Comparison of Detection Timeliness

FAR	MSJ (d_{msj})	S+CUSUM (d_{cu})	T+MA (d_{ma})	$d_{cu} - d_{msj}$	$d_{ma} - d_{msj}$
			Scenario 1		
0.0000	53.4	65.0	65.0	11.57 (0.44)	11.57 (0.44)
0.0125	30.2	52.8	22.7	22.57 (0.20)	-7.57 (0.54)
0.0250	27.0	50.0	20.1	23.07 (0.22)	-6.82 (0.52)
0.0500	22.6	42.6	15.2	20.04 (0.32)	-7.43 (0.52)
0.0750	21.4	37.4	12.8	16.00 (0.40)	-8.57 (0.46)
0.1000	18.6	32.9	10.6	14.36 (0.44)	-7.96 (0.36)
			Scenario 2		
0.000	13.4	28.0	27.6	14.64 (0.12)	14.22 (0.14)
0.012	7.1	25.7	6.9	18.60 (< 0.01)	-0.24 (0.96)
0.025	6.1	25.5	5.5	19.38 (< 0.01)	-0.64 (0.60)
0.050	5.2	25.3	4.5	20.16 (< 0.01)	-0.62 (0.62)
0.075	4.7	25.0	3.9	20.30 (< 0.01)	-0.86 (0.46)
0.100	4.7	25.0	3.7	20.32 (< 0.01)	-0.96 (0.44)

Table 6 summarizes the detection sensitivity of the surveillance methods considered. The first three columns report the sensitivity at different FARs. The last two columns report the differences of sensitivity between our approach and the benchmark methods. The parentheses in the last two columns are the p-values of the statistical tests hypothesizing equal performance of the two approaches. The p-value was computed using a bootstrap method based on the paired comparison of all testing days [68].

Because of the large number of testing days, it is not surprising to see significant testing results across all FARs. The p-values indicate that our approach is significantly better than the two benchmark methods across all FARs under consideration.

To better understand the intuition behind the performance difference, we present the alert scores around an outbreak period from the three surveillance methods. As plotted in Fig. 5a, the top panel is the input time series to the surveillance methods. The following three panels present the alert scores from our approach, the S+CUSUM method, and the T+MA method. The vertical solid lines in the three lower panels mark the beginning and ending of the outbreak period. The horizontal dashed lines mark the

thresholds corresponding to a FAR of 0.0125. The alert scores higher than the thresholds are marked as outbreak days by the three methods.

In this example, the outbreak lasted for 60 days. The MSJ method detected the outbreak first at the 28th day and continued through the 58th day. The T+MA method first detected the outbreak at the 20th day. However, the alert score fell below the threshold the next day and fluctuated around the threshold for the next few days. It was not until the 28th day when the alert scores move beyond the threshold steadily and fell below the threshold again at the 42nd day. The S+CUSUM method did not output scores higher than the threshold during the outbreak period. In total, our method detected 31 outbreak days, compared to 20 days by the T+MA method and 0 day by the S+CUSUM method.

Clearly, our approach had the best sensitivity. More than half of the outbreak days were correctly detected. Compared to our approach, the T+MA method detected the onset of the outbreak with similar speed. However, our approach did a much better job in detecting the ending of the outbreak. The ability to better detecting the ending of an outbreak contributes to the higher detection sensitivity of our approach.

One salient characteristic of this input time series is a jump at the end of the year "2109." Both the T+MA method and the S+CUSUM method were seriously disrupted by the jump, causing elevated alert scores. Our approach, on the other hand, filtered out the jump effectively and output reasonably low scores around the jump day. The unique ability to filter out the jumps keeps the false alarm rate low and leads to a better detection sensitivity.

Fig. 5b plots another example from Scenario 2. The simulated outbreak started from 8/24/1996 and lasted for 28 days. Note that all methods reported escalated alert scores at the beginning of year 1996, indicating a possible outbreak. Using the threshold associated with a FAR of 0.025, we can see that MSJ, T+MA, and S+CUSUM methods marked 17, 11, and 8 days as having outbreaks during late January to early February. From a domain perspective, it is extremely hard to verify whether there was actually an
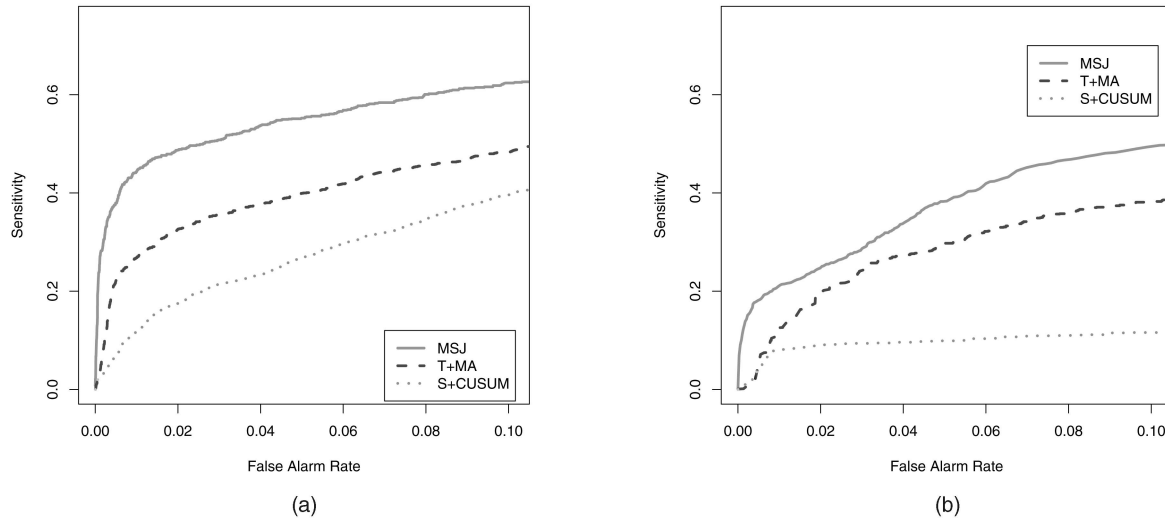

(a)

(b)

Fig. 4. Performance Comparison: Sensitivity. (a) Scenario 1 sensitivity at different false alarm rates. (b) Scenario 2 sensitivity at different false alarm rates.

TABLE 6
Comparison of Detection Sensitivity

FAR	MSJ (s_{msj})	S+CUSUM (s_{cu})	T+MA (s_{ma})	$s_{cu} - s_{msj}$	$s_{ma} - s_{msj}$
			Scenario 1		
0.0000	0.008	0.000	0.000	0.008 (< 0.01)	0.008 (< 0.01)
0.0125	0.459	0.140	0.286	0.320 (< 0.01)	0.174 (< 0.01)
0.0250	0.498	0.195	0.341	0.302 (< 0.01)	0.156 (< 0.01)
0.0500	0.552	0.267	0.399	0.286 (< 0.01)	0.153 (< 0.01)
0.0750	0.590	0.331	0.450	0.259 (< 0.01)	0.141 (< 0.01)
0.1000	0.624	0.395	0.484	0.228 (< 0.01)	0.140 (< 0.01)
			Scenario 2		
0.0000	0.070	0.000	0.001	0.070 (< 0.01)	0.069 (< 0.01)
0.0125	0.217	0.083	0.142	0.134 (< 0.01)	0.075 (< 0.01)
0.0250	0.266	0.093	0.216	0.173 (< 0.01)	0.050 (< 0.01)
0.0500	0.383	0.099	0.298	0.284 (< 0.01)	0.086 (< 0.01)
0.0750	0.461	0.109	0.355	0.352 (< 0.01)	0.106 (< 0.01)
0.1000	0.497	0.116	0.382	0.381 (< 0.01)	0.115 (< 0.01)

outbreak during this period and exactly how long the outbreak lasted. All three methods under consideration, nevertheless, were exposed to the same "noise" and thus the relative performance among the three methods should still be meaningful.

It is interesting to observe that even though the MSJ marked the most false positives during early 1996, it actually detected the highest number of outbreak days (7) compared to T+MV (4 days) and S+CUSUM (0 day) under a fixed FAR (0.025). In this particular example, the MSJ method detected the outbreak two days before T+MV. The worse performance of T+MV may be caused by the higher alert scores during early 1996 compared to those during the outbreak period. The MSJ method automatically adjusted the alert score to the range of 0 and 1 and thus was more sensitive and detected the outbreak earlier.

We also note that S+CUSUM seemed to overreact to the deviation during early 1996, and the alert scores during that period of time were almost twice as those generated by the simulated outbreak. It may be caused by the deterministic natural of the Serfling model. The estimated variance of the Serfling model was too small compared to that of the ARIMA model. Smaller variance leads to higher standardized errors and elevated alert scores.

5 CONCLUSION

Disease outbreak detection using time series data is an important function for syndromic surveillance systems. We treated the disease outbreak as hidden outbreak states and developed a Markov switching with jumps model for syndromic surveillance. To handle the negative effect caused by the jumps in the observed time series, we extended the Markov switching model to include an extreme value filtering component. The negative effect of jumps can be successfully filtered out, which led to a lower false alarm rate.

We evaluated our disease outbreak detection approach using both simulated and real-world baseline time series, together with outbreaks simulated following established methods. Two benchmark surveillance methods were included. The first benchmark method, S+CUSUM, uses the Serfling model to filter out seasonal fluctuation and then applies the CUSUM method on standardized prediction errors. The second benchmark method, T+MA, uses trimmed-mean seasonal ARMA model and computes the alert scores using linear-increasing weights. The evaluation results showed that our method achieved a similar level of detection timeliness and higher detection sensitivity compared to the benchmark outbreak detection methods. Our approach had a detection sensitivity 23-328 percent higher than the benchmark methods.
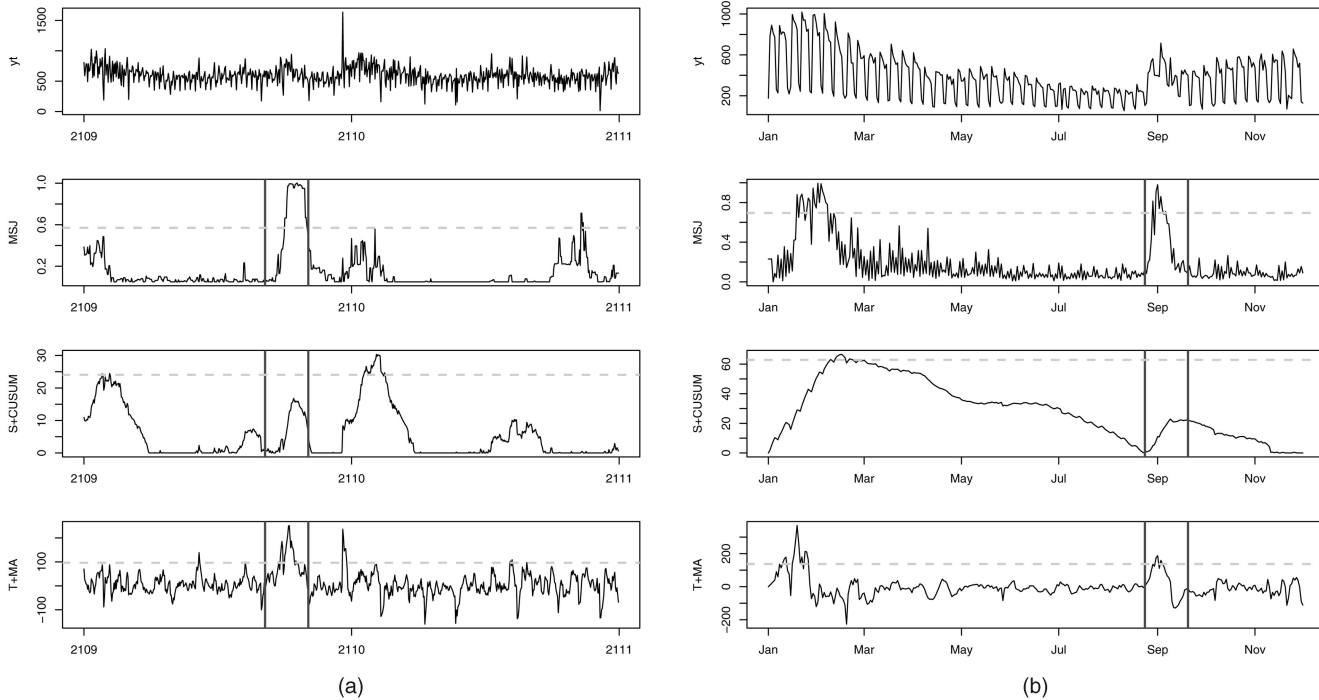

Fig. 5. A comparison of the alert scores from three surveillance methods. (a) Prospective Detection Details: A case from Scenario 1. (b) Prospective Detection Details: A case from Scenario 2 (year 1996).

We applied an earlier version of the detection methods reported in this paper in a disease outbreak detection algorithm competition organized by the International Society for Disease Surveillance in 2007. Our method ranked third among participating methods. The performance gap between our method and the best performing algorithm is within 6 percent.

The results reported in our study suggest a promising future for the use of hidden state variables to model the changing dynamics of observed surveillance time series. We plan to extend our approach to outbreak detection with multiple data streams through multivariate time series analysis based on Markov switching. We are also exploring opportunities to apply the approach developed in this paper in areas beyond infectious disease informatics. One such area is sensor data integration and anomaly detection.

REFERENCES

[1] P.-H. Hu, D. Zeng, H. Chen, C. Larson, W. Chang, C. Tseng, and J. Ma, "System for Infectious Disease Information Sharing and Analysis: Design and Evaluation," *IEEE Trans. Information Technology in Biomedicine,* vol. 11, no. 4, pp. 483-492, July 2007.

[2] S. Niiranen, J. Yli-Hietanen, and L. Nathanson, "Toward Reflective Management of Emergency Department Chief Complaint Information," *IEEE Trans. Information Technology in Biomedicine,* vol. 12, no. 6, pp. 763-767, Nov. 2008.

[3] W.W. Chapman, L.M. Christensen, M.M. Wagner, P.J. Haug, O. Ivanov, J.N. Dowling, and R.T. Olszewski, "Classifying Free-Text Triage Chief Complaints into Syndromic Categories with Natural Language Processing," *Artificial Intelligence in Medicine,* vol. 33, no. 1, pp. 31-40, 2005.

[4] O. Ivanov, M.M. Wagner, W.W. Chapman, and R.T. Olszewski, "Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance," *Proc. Am. Medical Informatics Assoc. (AMIA) Symp.,* pp. 345-349, 2002.

[5] P. Yan, H. Chen, and D. Zeng, "Syndromic Surveillance Systems," *Ann. Rev. of Information Science and Technology,* vol. 42, pp. 425-495, 2008.

[6] J. Espino, M. Wagner, F. Tsui, H. Su, R. Olszewski, Z. Lie, W. Chapman, X. Zeng, L. Ma, Z. Lu, and J. Dara, "The RODS Open Source Project: Removing a Barrier to Syndromic Surveillance," *Studies in Health Technology and Informatics,* vol. 107, pp. 1192-1196, 2004.

[7] K.D. Mandl, M. Overhage, M. Wagner, W. Lober, P. Sebastiani, F. Mostashari, J. Pavlin, P.H. Gesteland, T. Treadwell, E. Koski, L. Hutwagner, D.L. Buckeridge, R.D. Aller, and S. Grannis, "Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience," *J. Am. Medical Informatics Assoc.,* vol. 11, no. 2, pp. 141-150, 2004.

[8] B.Y. Reis, M. Pagano, and K.D. Mandl, "Using Temporal Context to Improve Biosurveillance," *Proc. Nat'l Academy of Sciences USA,* vol. 100, pp. 1961-1965, Feb. 2003.

[9] B.Y. Reis and K.D. Mandl, "Time Series Modeling for Syndromic Surveillance," *BMC Medical Informatics and Decision Making,* vol. 3, no. 2, Jan. 2003.

[10] J. Takeuchi and K. Yamanishi, "A Unifying Framework for Detecting Outliers and Change Points from Time Series," *IEEE Trans. Knowledge and Data Eng.,* vol. 18, no. 4, pp. 482-492, Apr. 2006.

[11] W.A. Shewhart, *Statistical Method from the Viewpoint of Quality Control.* Department of Agriculture, The Graduate School, 1939.

[12] D.C. Montgomery, *Introduction to Statistical Quality Control,* fifth ed. Wiley, 2005.

[13] E.S. Page, "Continuous Inspection Schemes," *Biometrika,* vol. 41, nos. 1/2, pp. 100-115, June 1954.

[14] CDC, "Increased Antiviral Medication Sales Before the 2005-06 Influenza Season-New York City," *Morbidity and Mortality Weekly Report,* vol. 55, pp. 277-279, Mar. 2006.

[15] D.L. Buckeridge, P. Switzer, D. Owens, D. Siegrist, J. Pavlin, and M. Musen, "An Evaluation Model for Syndromic Surveillance: Assessing the Performance of a Temporal Algorithm," *Morbidity and Mortality Weekly Report,* vol. 54, pp. 109-115, Aug. 2005.

[16] M.P. Clements and D.F. Hendry, "Forecasting with Breaks," *Handbook of Economic Forecasting,* vol. 1, pp. 605-657, Elsevier, 2006.

[17] C.-S.J. Chu, M. Stinchcombe, and H. White, "Monitoring Structural Change," *Econometrica,* vol. 64, no. 5, pp. 1045-1065, 1996.

[18] J.D. Hamilton, "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica,* vol. 57, no. 2, pp. 357-84, Mar. 1989.

[19] B.Y. Reis and K.D. Mandl, "Integrating Syndromic Surveillance Data Across Multiple Locations: Effects on Outbreak Detection Performance," *Proc. Am. Medical Informatics Assoc. (AMIA) 2003 Symp.,* pp. 549-553, 2003.

[20] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control.* Holden Day, 1970.

[21] W.H. Greene, *Econometric Analysis.* Prentice Hall, 2000.

[22] H. Akaike, "Statistical Predictor Identification," *Annals of the Inst. of Statistical Math.,* vol. 22, pp. 203-217, 1970.

[23] H. Akaike, "Information Theory and an Extension of the Likelihood Principle," *Proc. Second Int'l Symp. Information Theory,* B.N. Perov and F. Csaki, eds., 1973.

[24] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 461-464, 1978.

[25] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[26] H. White, *Approximate Nonlinear Forecasting Methods,* vol. 1, chapter 9, pp. 459-512. Elsevier, Jan. 2006.

[27] J. Shao, "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica,* vol. 7, pp. 221-264, 1997.

[28] M.L. Jackson, A. Baer, I. Painter, and J. Duchin, "A Simulation Study Comparing Aberration Detection Algorithms for Syndromic Surveillance," *BMC Medical Informatics and Decision Making,* vol. 7, no. 6, 2007.

[29] S.C. Wieland, J.S. Brownstein, B. Berger, and K.D. Mandl, "Automated Real Time Constant-Specificity Surveillance for Disease Outbreaks," *BMC Medical Informatics and Decision Making,* vol. 7, no. 15, 2007.

[30] J. Zhang, F.-C. Tsui, M.M.W. William, and R. Hogan, "Detection of Outbreaks from Time Series Data Using Wavelet Transform," *Proc. Am. Medical Informatics Assoc. (AMIA) Symp.,* 2003.

[31] R. Serfling, "Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths," *Public Health Reports,* vol. 78, pp. 494-506, 1963.

[32] J.C. Brillman, T. Burr, D. Forslund, E. Joyce, R. Picard, and E. Umland, "Modeling Emergency Department Visit Patterns for Infectious Disease Complaints: Results and Application to Disease Surveillance," *BMC Medical Informatics and Decision Making,* vol. 5, no. 4, 2005.

[33] C.C. Holt, "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages," *Int'l J. Forecasting,* vol. 20, no. 1, pp. 5-10, 2004.

[34] P.R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science,* vol. 6, no. 3, pp. 324-342, 1960.

[35] H.S. Burkom, S.P. Murphy, and G. Shmueli, "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine,* vol. 26, pp. 4202-4218, Sept. 2007.

[36] J. Hamilton, *Time Series Analysis.* Princeton, 1994.

[37] A.N. Shiryaev, "On Optimum Methods in Quickest Detection Problems," *Theory of Probability and Its Applications,* vol. 8, pp. 22-46, 1963.

[38] S.W. Roberts, "A Comparison of Some Control Chart Procedures," *Technometrics,* vol. 8, pp. 411-430, 1966.

[39] M. Frisen and J. De Mare, "Optimal Surveillance," *Biometrika,* vol. 78, no. 2, pp. 271-280, 1991.

[40] C. Sonesson and D. Book, "Review and Discussion of Prospective Statistical Surveillance in Public Health," *J. Royal Statistical Soc., Series A,* vol. 166, no. 1, pp. 5-21, 2003.

[41] G.V. Moustakides, "Optimal Stopping Times for Detecting Changes in Distributions," *Annals of Statistics,* vol. 14, no. 4, pp. 1379-1387, Dec. 1986.

[42] M. Frisen, "Statistical Surveillance. Optimality and Methods," *Int'l Statistical Rev.,* vol. 71, no. 2, pp. 403-434, 2003.

[43] S. Chandrasekaran, J.R. English, and R.L. Disney, "Modeling and Analysis of EWMA Control Schemes with Variance-Adjusted Control Limits," *IIE Trans.,* vol. 27, pp. 282-290, 1995.

[44] S.H. Steiner, "EWMA Control Charts with Time-Varying Control Limits and Fast Initial Response," *J. Quality Technology,* vol. 31, no. 1, pp. 75-86, 1999.

[45] C. Sonesson, "Evaluations of Some Exponentially Weighted Moving Average Methods," *J. Applied Statistics,* vol. 30, no. 10, pp. 1115-1133, 2003.

[46] H. Burkom, "Alerting Algorithms for Biosurveillance," *Disease Surveillance: A Public Health Informatics Approach,* pp. 143-192, John Wiley & Sons, 2007.

[47] C.-J. Kim and C.R. Nelson, *State-Space Models with Regime Switching.* MIT Press, 1999.

[48] L.E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Annals of Math. Statistics,* vol. 37, pp. 1554-1563, 1966.

[49] L.E. Baum and J.A. Egon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," *Bull. Am. Meteorology Soc.,* vol. 73, pp. 360-363, 1967.

[50] Y.L. Strat and F. Carrat, "Monitoring Epidemiologic Surveillance Data Using Hidden Markov Models," *Statistics in Medicine,* vol. 18, pp. 3463-3478, 1999.

[51] M. Dahlquist and S.F. Gray, "Regime-Switching and Interest Rates in the European Monetary System," *J. Int'l Economics,* vol. 50, no. 2, pp. 399-419, Apr. 2000.

[52] S.L. Scott, "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century," *J. Am. Statistical Assoc.,* vol. 97, pp. 337-351, 2002.

[53] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc., Series B (Methodological),* vol. 39, no. 1, pp. 1-38, 1977.

[54] C.A. Popescu and Y.S. Wong, "Nested Monte Carlo EM Algorithm for Switching State-Space Models," *IEEE Trans. Knowledge and Data Eng.,* vol. 17, no. 12, pp. 1653-1663, Dec. 2005.

[55] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional Anomaly Detection," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 5, pp. 631-645, May 2007.

[56] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *Am. Statistician,* vol. 49, no. 4, pp. 327-335, Nov. 1995.

[57] J.H. Albert and S. Chib, "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *J. Business & Economic Statistics,* vol. 11, no. 1, pp. 1-15, Jan. 1993.

[58] C.K. Carter and R. Kohn, "On Gibbs Sampling for State Space Models," *Biometrika,* vol. 81, no. 3, pp. 541-553, Aug. 1994.

[59] D. Madigan, "Bayesian Data Mining for Health Surveillance," *Spatial and Syndromic Surveillance for Public Health,* pp. 203-221, John Wiley & Sons, 2005.

[60] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems," *J. Royal Statistical Soc., Series B (Methodological),* vol. 36, no. 2, pp. 192-236, 1974.

[61] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 6, pp. 721-741, Nov. 1984.

[62] ISDS, "My Algorithm Can Out-Detect Your Algorithm: Biosurveillance Using Time Series Data," technical report, Int'l Soc. for Disease Surveillance, https://wiki.cirg.washington.edu/pub/bin/view/Isds/TechnicalContest; accessed, Nov. 2008.

[63] L.M. Wein, D.L. Craft, and E.H. Kaplan, "Emergency Response to an Anthrax Attack," *Proc. Nat'l Academy of Sciences USA,* vol. 100, pp. 4346-4351, Apr. 2003.

[64] J.A. Jernigan, D.S. Stephens, D.A. Ashford, C. Omenaca, M.S. Topiel, M. Galbraith, M. Tapper, T.L. Fisk, S. Zaki, T. Popovic, R.F. Meyer, C.P. Quinn, S.A. Harper, S.K. Fridkin, J.J. Sejvar, C.W. Shepard, M. McConnell, J. Guarner, W.J. Shieh, J. Malecki, J.L. Gerberding, J.M. Hughes, and B.A. Perkins, "Bioterrorism-Related Inhalational Anthrax: The First 10 Cases Reported in the United States," *Emerging Infectious Diseases,* vol. 7, pp. 933-944, 2001.

[65] T. Burr, T. Graves, R. Klamann, S. Michalak, R. Picard, and N. Hengartner, "Accounting for Seasonal Patterns in Syndromic Surveillance Data for Outbreak Detection," *BMC Medical Informatics and Decision Making,* vol. 6, no. 40, 2006.

[66] H. Rolka, H. Burkom, G.F. Cooper, M. Kulldorff, D. Madigan, and W.-K. Wong, "Issues in Applied Statistics for Public Health Bioterrorism Surveillance Using Multiple Data Stream: Research Needs," *Statistics in Medicine,* vol. 26, pp. 1834-1856, 2007.

[67] B. Efron and R. Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science,* vol. 1, no. 1, pp. 54-75, 1986.

[68] H.-M. Lu, D. Zeng, L. Trujillo, K. Komatsu, and H. Chen, "Ontology-Enhanced Automatic Chief Complaint Classification for Syndromic Surveillance," *J. Biomedical Informatics,* vol. 41, no. 2, pp. 340-356, Apr. 2008.

Hsin-Min Lu received the bachelor's degree in business administration from the National Taiwan University, Taipei, in 1998, and the MA degree in economics from the National Taiwan University, Taipei, in 2002. He is currently working toward the PhD degree in the Management Information Systems Department, University of Arizona, Tucson. His research interests include data mining, text mining, and applied econometrics.

Daniel Zeng received the MS and PhD degrees in industrial administration from Carnegie Mellon University, Pittsburgh, Pennsylvania. He is an associate professor and a Honeywell fellow in the Department of Management Information Systems at the University of Arizona, Tucson, and a research professor at the Institute of Automation, the Chinese Academy of Sciences. His research interests include software agents and their applications, security informatics, social computing, computational game theory, recommender systems, and spatiotemporal data analysis. He has coedited 15 books and published more than 100 peer-reviewed articles in information systems and computer science journals, edited books, and conference proceedings. He is a senior member of the IEEE.

Hsinchun Chen received the BS degree from the National Chiao-Tung University in Taiwan, the MBA degree from the State University of New York at Buffalo, and the PhD degree in information systems from New York University. He is a McClelland professor of management information systems at the University of Arizona. He has served as a scientific counselor/advisor of the US National Library of Medicine, the Academia Sinica (Taiwan), and the National Library of China (China). He is a fellow of the IEEE and the AAAS. He received the IEEE Computer Society 2006 Technical Achievement Award. He was ranked #8 in publication productivity in information systems (CAIS 2005) and #1 in Digital Library research (IP&M 2005) in two bibliometric studies. His COPLINK system, which has been quoted as a national model for public safety information sharing and analysis, has been adopted in more than 550 law enforcement and intelligence agencies in 20 states.