NISS

A Spatio-Temporal Absorbing State Model for Disease and Syndromic Surveillance

Matthew J. Heaton, Frank Zou,
David L. Banks, Alan F. Karr,
Gauri Datta, James Lynch and Francisco Vera

Technical Report 175
October, 2010

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC
www.niss.org

A Spatio-Temporal Absorbing State Model for Disease and Syndromic Surveillance

Matthew J. Heaton[1,*], Frank Zou[2], David L. Banks[1], Alan Karr[2],
Gauri Datta[3], James Lynch[4], and Francisco Vera[5]

[1]*Department of Statistical Science, Duke University, Box 90251, Durham NC 27708-025*
[2]*National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park NC 27709-4006*
[3]*Department of Statistics, University of Georgia, Athens GA 30602-7952*
[4]*Department of Statistics, University of South Carolina, Columbia SC 29208*
[5]*Escuela Superior Politecnica del Litoral, Ecuador*

SUMMARY

Reliable surveillance models are an important tool in public health because they aid in mitigating disease outbreaks, identify where and when disease outbreaks occur, and predict future occurrences. While many statistical models have been devised for surveillance purposes, none are able to simultaneously achieve important practical goals such as good sensitivity and specificity, proper use of domain information, inclusion of spatio-temporal dynamics, and transparent support to decision-makers. In an effort to achieve some of these goals, this paper proposes a spatio-temporal conditional autoregressive hidden Markov model with an absorbing state. The model performs well in both a large simulation study and in an application to influenza/pneumonia fatality data.

KEY WORDS: Conditional autoregressive model; Hierarchical model; Hidden Markov model; Influenza.

* Corresponding Author: Matthew J. Heaton, email: matt@stat.duke.edu, phone: 919-684-4558, fax: 919-684-8594

1. INTRODUCTION

The goal of disease and syndromic surveillance is to monitor and detect aberrations in disease prevalence across space and time. Disease surveillance typically refers to the monitoring of confirmed cases of disease while syndromic surveillance uses syndromes associated with disease to detect aberrations. In either situation, any proper surveillance system should be able to (i) detect, as early as possible, potentially harmful deviations from baseline levels of disease while maintaining low false positive detection rates, (ii) incorporate the spatial and temporal dynamics of a disease system, (iii) be widely applicable to multiple diseases or syndromes, (iv) incorporate covariate information, and (v) produce results which are readily interpretable by policy decision makers.

The literature describes many methods, algorithms, and models designed to support surveillance systems; it spans multiple disciplines including statistics, computer science, epidemiology, and public health. Early approaches to surveillance were primarily computational algorithms. For example, the CUSUM [1] technique and its variants [see, e.g., 2, 3, 4] monitor the cumulative deviation (over time) of disease counts from some baseline rate. Qiu and Hawkins [5] and Mason et al. [6] adapt the CUSUM technique for monitoring multiple diseases. While CUSUM techniques have been popular, they are quite sensitive to the baseline level and are difficult to adapt to situations in which there is spatial correlation.

A second line of work uses spatial scan statistics, originally proposed by Kulldorff [7] with later extensions given in Walther [8], Neill and Cooper [9], Kulldorff et al. [10], and Neill et al. [11]. Scan statistics are designed to find spatial clusters where disease occurrences are high. However, these algorithms are often computationally expensive, do not easily account for covariate information, and fail to provide an easily interpretable measure of uncertainty associated with the identified cluster.

The majority of recent surveillance techniques are model-based approaches. Advantages of model-based approaches include the ability to incorporate covariate information, the flexibility in accounting for spatial and temporal dynamics in a hierarchical framework, and direct interpretability of model parameters which facilitates decision-making. For example, LeStrat and Carrat [12] pioneered the use of hidden Markov models (HMMs) for use in surveillance applications. Because of their work, HMM based surveillance methods have seen widespread use—an overview of such models is provided in Madigan [13]. Martínez-Beneito et al. [14] and Rath et al. [15] detail model-based methods for detecting influenza outbreaks in a purely temporal setting (i.e., they use a time series analysis but do not account for spatial structure). Knorr-Held and Richardson [16] discuss a spatio-temporal HMM for meningococcal disease incidence; among the previous research, their approach is the most similar to the method developed in this paper.

The contribution of this article is to improve model-based surveillance methods by (i) detailing the structure of a hierarchical HMM for the surveillance of disease across space and time and (ii) proposing a new, non-separable spatio-temporal autoregressive model. Because surveillance data is typically gathered at discrete time periods (such as weeks, days, etc.) and at a discrete number of spatial locations (e.g. hospitals, counties, etc.), the focus of this paper is on models that apply to such situations—models for continuous space/time are left for future work. Additionally, the focus of this article is on surveillance of a single disease or syndrome.

One novelty introduced here is the use of an absorbing state Markov chain to model the epidemic state. By considering the epidemic state to be absorbing, undesirable behavior such

as day-to-day switching between epidemic and non-epidemic states is avoided. Forcing non-switching behavior is reasonable in that if an outbreak of a disease occurs at time t, then the outbreak should still be occurring at time $t + 1$. Of course, by using an absorbing state Markov chain, the model is limited in application to a *single* outbreak of a disease at each location—this is appropriate when the goal is to quickly discover the emergency (as in bioterrorism or a pandemic). On the other hand, the methods described herein are not designed to monitor influenza over several years because multiple outbreaks of the flu will have occurred within the observed time period.

Beyond the use of an absorbing state, the second contribution of this article is to allow the non-absorbing transition probabilities of the Markov chain to vary over space and time. Specifically, the probability of a spatial location transitioning from a non-epidemic state to an epidemic one is assumed to be conditional on the state of its neighbors at the previous time period and (potentially) relevant covariates such as population size, age distribution, and so forth. Specifically, the probability of transitioning from a non-epidemic state at time $t-1$ to an epidemic one at time t depends on the number of neighbors in an epidemic state at time $t - 1$. Such models appropriately capture the spatial and temporal dynamics of disease propagation within a spatio-temporal network.

Section 2 presents a general hierarchical framework for a surveillance model, including interpretation of model parameters in the surveillance context and details of an absorbing state surveillance model. Section 3 discusses model inference. Section 4 evaluates the sensitivity of the model as a surveillance tool to various parameter specifications. The model is then applied to a surveillance data set of deaths due to influenza and pneumonia across 121 cities in the United States during a single flu season. Section 6 summarizes the findings of this article and provides directions for future research.

2. SURVEILLANCE MODELS

2.1 Infrastructure of a Surveillance Model

As a general surveillance strategy, the basic model described in Banks et al. [17] is adopted here. Specifically, let $Y_s(t)$ denote a univariate disease or syndrome count for spatial location $s = 1, \ldots, S$ at time $t = 1, \ldots, T$. As the data layer for a hierarchical HMM, let

$$Y_s(t) \sim \mathcal{P}(\mu_s(t) + \delta_s(t)\lambda_s(t)) \tag{1}$$

and assume the $\{Y_s(t) \; \forall \; s, t\}$ are independent given the parameters $\mu_s(t), \delta_s(t)$, and $\lambda_s(t)$. In the surveillance context, $\mu_s(t) > 0$ represents a baseline rate of disease during a non-epidemic stage and $\delta_s(t) \in \{0, 1\}$ is an indicator where $\delta_s(t) = 1$ implies that an epidemic is occurring at location s at time t. The parameter $\lambda_s(t) > 0$ is the expected additive increase of disease counts due to the epidemic. Notice that $\kappa_s(t) \equiv \lambda_s(t)/\mu_s(t) > 0$ quantifies the proportional increase in case counts during the epidemic period relative to the non-epidemic period. When $\delta_s(t) = 0$ the parameter $\lambda_s(t)$ is not identified as it will not contribute to the likelihood; hence, assuming $\lambda_s(t) \equiv 0$ when $\delta_s(t) = 0$ is necessary for model identifiability.

The use of the Poisson likelihood in (1) is the most natural choice because it describes integer valued random variables. However, for large values of $Y_s(t)$, a Gaussian approximation (perhaps

based on a log-transformation) is also a suitable likelihood and may lead to more efficient parameter estimation. Furthermore, over-dispersion can be incorporated by assuming a negative binomial distribution as in [18] and Held et al. [19]. Since our primary focus is on the detection of outbreaks early in the epidemic, then small disease counts are to be expected; therefore (1) is an appropriate likelihood specification.

For the baseline rate, let

$$\log(\mu_s(t)) \equiv \alpha_\mu + \boldsymbol{x}_s'(t)\boldsymbol{\beta}_\mu + \xi_s(t), \tag{2}$$

where α_μ is a global baseline rate, $\boldsymbol{x}_s(t) = (x_{s,1}(t), \ldots, x_{s,p}(t))'$ is a vector of p covariates, $\boldsymbol{\beta}_\mu = (\beta_{\mu,1}, \ldots, \beta_{\mu,p})$ is the associated vector of coefficients, and $\xi_s(t)$ is a spatio-temporal random effect. The $\boldsymbol{x}_s(t)$ contains information describing location s at time t, such as population density, age distribution, income level, and so forth.

In the presence of informative covariates, a simplifying assumption for the baseline rate is that the covariate information accounts for all residual spatial and temporal correlation such that $\xi_s(t) \equiv 0$ for all s and t. However, if an application lacks essential covariate information this assumption will be inappropriate, necessitating a spatio-temporal model for $\xi_s(t)$. Knorr-Held and Richardson [16] assume a separable space-time model such that $\xi_s(t) = \psi_s + \gamma_t$ where ψ_s follows an intrinsic autoregressive model [20, 21 sec. 3.3] and γ_t is a temporal term accounting for seasonal and other temporal correlations. While a separable space-time effect may be appropriate for some applications, more general space-time models are needed. For example, contagious disease processes motivate the non-separable space-time model presented in Section 2.2.

The indicators $\{\delta_s(t)\}$ are typically modeled as a two-state Markov chain with transition matrix $\boldsymbol{\Gamma}_s(t) = \{\gamma_{s,ij}(t)\}$ where $\gamma_{s,ij}(t)$ represents the probability that, starting in state i, location s transitions to state j in the interval from time t to $t+1$. The common simplifying assumption is that $\boldsymbol{\Gamma}_s(t) \equiv \boldsymbol{\Gamma}$ for all s and t but this assumption is inappropriate when surveilling diseases over a large spatial region and many time periods (and thus this assumption is not used here). Furthermore, the support of $\delta_s(t)$ need not be $\{0, 1\}$ (although this is commonly assumed). For example, Mugglin et al. [22] used a k-state Markov chain to describe various stages of an epidemic; however, a k-state model is tuned to describe stages of an epidemic as opposed to detecting the first outbreak.

The parameter $\lambda_s(t)$ represents the additive increase due to the presence of an epidemic. As a general model for $\lambda_s(t)$, we assume a similar structure to $\mu_s(t)$, so

$$\log(\lambda_s(t)) = \alpha_\lambda + \boldsymbol{h}_s'(t)\boldsymbol{\beta}_\lambda + \theta_s(t), \tag{3}$$

where α_λ is the global increase, $\boldsymbol{h}_s(t) = (h_{s,1}(t), \ldots, h_{s,m}(t))'$ is a vector of covariates with associated coefficients $\boldsymbol{\beta}_\lambda$, and $\theta_s(t)$ is the spatio-temporal random effect. Most applications will either assume $\boldsymbol{h}_s(t) = \boldsymbol{x}_s(t)$ in which case β_λ represents the additive effect of each covariate in an epidemic state, or else $\boldsymbol{h}_s(t) = 0$. Perhaps in some special cases $\boldsymbol{h}_s(t)$ can contain some information on the disease or syndrome being surveilled but such cases are rare. The spatio-temporal structure of $\theta_s(t)$ is bound to be more complex than that of $\xi_s(t)$ due to complex social and spatial networks when a disease is present. For example, in an epidemic state, Los Angeles and New York City could be considered spatial neighbors due to air transportation but in a non-epidemic state these two regions are expected to behave differently due to spatial location.

2.2 An Absorbing State Model

With the infrastructure for a spatio-temporal surveillance model described in Section 2.1 in place, this section describes an absorbing state model and a few novel modeling strategies. First, assume $\xi_s(t) \equiv \xi_s$ for all t so that the baseline rate of disease for spatial location s in a non-epidemic period is $\exp\{\alpha_\mu + \xi_s\}$. As such, any day-to-day or week-to-week variation in disease counts is assumed to be attributable to chance variation. As a model for ξ_s, let ξ_s follow an intrinsic autoregressive model [23] such that

$$\xi_s \mid \xi_{i \neq s} \sim \mathcal{N}\left(\frac{1}{w_{\xi,+s}} \sum_{i \neq s} w_{\xi,is}\xi_i, \frac{\sigma_\xi^2}{w_{\xi,+s}}\right) \tag{4}$$

where $w_{\xi,is}$ is the spatial weight of location i on location s during a non-epidemic period and $w_{\xi,+s} = \sum_{i \neq s} w_{\xi,is}$. And let

$$w_{\xi,is} = \begin{cases} 1 & \text{if } i \text{ and } s \text{ share a border,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This simple spatial model for ξ_s is appropriate in this setting because spatial neighbors should exhibit similar baseline rates of disease. As mentioned in Banerjee et al. [21], the intrinsic autoregressive model is improper in the sense that the joint distribution $[\xi_1, \ldots, \xi_S]$ is improper. To ensure propriety, the $\{\xi_s\}$ are constrained such that $\sum_s \xi_s = 0$. With this constraint, α_μ provides the centering of the baseline disease rate.

As mentioned previously, the spatio-temporal process for $\theta_s(t)$ should be more complicated than that of ξ_s. Indeed, the assumption that $\theta_s(t) \equiv \theta_s$ for all t is incorrect for an epidemic period because once a disease is present, counts will certainly trend upward over time and high disease counts will tend to persist. As such, assume $\theta_s(t)$ follows a non-separable space time model similar to that described in Banks et al. [17], so

$$\theta_s(t) \mid \theta_{i \neq s}(t), \text{past} \sim \mathcal{N}\left(\rho_\theta \theta_s(t-1) + \frac{1}{w_{\theta,+s}} \sum_{i \neq s} w_{\theta,is}\theta_i(t), \frac{\sigma_\theta^2}{w_{\theta,+s}+1}\right) \tag{6}$$

where $w_{\theta,is}$ is a spatial weight of location i on location s during an epidemic state and $w_{\theta,+s} = \sum_i w_{\theta,is}$. The parameter $\rho_\theta \in (0,1)$ describes the dependence of disease counts at time t on disease counts at time $t-1$ with higher values indicating a higher degree of dependence. The spatial process is again taken to be an intrinsic autoregressive process. Propriety for the non-separable spatio-temporal model given by (6) is ensured by the constraint that $\sum_s \theta_s(t) = 0$ for all t; in this case, α_λ provides the centering. The variance term σ_θ^2 is scaled by a factor $w_{\theta,+s}+1$ to reflect the information about $\theta_s(t)$ given by its $w_{\theta,+s}+1$ neighbors (i.e., $w_{\theta,+s}$ spatial neighbors and one temporal neighbor, itself in the previous time period).

The main interest here is in the early detection of epidemics. In other words, the goal here is to develop a model to detect the *first* outbreak as opposed to modeling the outbreak itself or subsequent waves of outbreaks in the same location. Hence, the purpose of the model has been realized once an outbreak has been detected. As such, a model which frequently switches between an epidemic and a non-epidemic state is undesirable, and it can be shown that if a

model is allowed to switch, it generates a large number of false alarms. To avoid frequent state switching, $\delta_s(t)$ is assumed to be a realization of an absorbing state Markov chain; i.e.,

$$\mathbb{P}\text{r}(\delta_s(t+1)=1 \mid \boldsymbol{\delta}(t)) = \begin{cases} \gamma_s(t) = \pi_1 \mathbb{I}_{\{\delta_i(t)=0 \,\forall\, i \in \mathcal{N}_s\}} + 1 - (1-\pi_2)^{\sum_{i\in\mathcal{N}_s}\delta_i(t)} & \text{if } \delta_s(t)=0, \\ 1 & \text{if } \delta_s(t)=1, \end{cases} \tag{7}$$

where $\boldsymbol{\delta}(t) = (\delta_1(1), \ldots, \delta_S(t))$ is a vector containing all $\delta_s(t)$ up to time t, $\mathbb{I}_{\mathcal{A}}$ is an indicator for the set $\mathcal{A}$, $\pi_1 \in (0,1)$ is a spontaneous disease generation rate, $\pi_2 \in (0,1)$ is a contagion rate, and $\mathcal{N}_s$ is a set of spatial and/or social network neighbors of location s. Notice that if the disease is absent at all neighbors of s at time t, a disease spontaneously arises at time $t+1$ with probability π_1. Alternatively, if a disease is present at neighbors of s (but not at location s) then the probability of location s becoming infected at time $t+1$ is given by $1-(1-\pi_2)^{\sum_{i\in\mathcal{N}_s}\delta_i(t)}$. This probability model is chosen to reflect the intuition for contagious diseases that the probability of an outbreak is an increasing function in (i) the number of neighbors of s infected at time t and (ii) π_2. Thus, if $\pi_2 \approx 1$ then the disease under surveillance has a high rate of spread.

The specification for transitions from non-epidemic to epidemic states given by (7) is novel in a few respects. First, the transition probabilities $\gamma_s(t)$ vary by spatial location and time. Previous work by, among others, Martínez-Beneito et al. [14] and Knorr-Held and Richardson [16] assume time and space homogenous transition probabilities. Second, by assuming an absorbing epidemic state, the model will not fluctuate rapidly between states.

3. MODEL INFERENCE

To complete the hierarchical specification and to perform model inference, prior distributions are required for the model parameters α_μ, α_λ, $\boldsymbol{\beta}_\mu$, $\boldsymbol{\beta}_\lambda$, σ^2_ξ, σ^2_θ, ρ_θ, π_1, and π_2. Independent Gaussian prior distributions are convenient prior distributions for α_μ, α_λ, $\boldsymbol{\beta}_\mu$, and $\boldsymbol{\beta}_\lambda$. The initial temptation is to let these Gaussian prior distributions be non-informative or even improper, but recall that each of these parameters is defined on the log scale of $\mu_s(t)$ and $\lambda_s(t)$. Hence, informative Gaussian distributions such as $\mathcal{N}(0,10)$ are relatively non-informative on the count scale. Assuming σ^2_ξ and σ^2_μ follow independent inverse-gamma prior distributions is a convenient choice because these lead to closed form complete conditional distributions from which samples can be drawn directly. A natural prior distribution for ρ_θ does not exist. Simulations (not shown) indicate that ρ_θ is difficult to estimate but that results are quite insensitive to its value. Hence, either a discrete prior or fixing the value of ρ_θ is most appropriate. For the simulations and applications performed here, the value of ρ_θ is fixed at 0.5. Finally, because $\pi_i \in (0,1)$ for $i=1,2$, the natural prior is the beta distribution. Generally, π_1 and π_2 are expected to be small; hence, the $\mathcal{B}e(1,30)$ distribution is used here.

Given the prior distributions above, model inference is done via Markov chain Monte Carlo (MCMC) simulation. Unfortunately, the majority of complete conditional distributions are not available in closed form and thus Metropolis-Hastings updates are needed. Because $\delta_s(t)$ is discrete, a forward filtering backward smoothing (FFBS) algorithm [see 24 sec. 2.5] can be implemented to sample $\boldsymbol{\delta}_s = (\delta_s(1), \ldots, \delta_s(T))'$ directly, which greatly improves mixing. A description of the FFBS algorithm is included in the Appendix. Admittedly, the use of MCMC for

model inference requires expert input to ensure proper mixing and convergence. However, recent work on "black box" implementations such as adaptive MCMC [see 25] or integrated nested Laplace approximations [see 26] show promise that such methods can soon be implemented and used by non-statisticians in the public health community.

For prospective analyses (i.e. analyses with the goal of detecting the current disease state), the main quantity of interest is the posterior distribution $[\boldsymbol{\delta}(T) \mid \boldsymbol{Y}]$ where $\boldsymbol{\delta}(T) = (\delta_1(T), \ldots, \delta_S(T))'$, $\boldsymbol{Y}$ is all the observed data up to time T, and $[\cdot]$ is a general probability distribution. However, a retrospective analysis of $[\delta_s(t) \mid \boldsymbol{Y}]$ can provide useful forensic insight on where and when the disease was likely to have been introduced. Regardless of whether the prospective or retrospective approach is adopted, a decision rule needs to be constructed such that time periods and spatial locations are determined to be in an epidemic state or not. Given the application of the model, a full decision theoretic treatment including costs of false-positives (declaring a disease to be present when it is not) and false-negatives (declaring a disease to be absent when it is not) is necessary in forming this decision rule. However, as a starting point, a convenient decision rule is based on the posterior mean of $[\delta_s(t) \mid \boldsymbol{Y}]$ because $\mathbb{E}(\delta_s(t) \mid \boldsymbol{Y}) = \Pr(\delta_s(t) = 1 \mid \boldsymbol{Y})$ and the posterior mean is the Bayes estimator under squared error loss. Selection of the decision rule is investigated further in Section 4.

4. SIMULATIONS

The effectiveness of the absorbing state model described in Section 2.2 as a surveillance tool is dependent on several factors. To name a few, diseases with low epidemic-to-baseline ratios ($\kappa_s(t) = \lambda_s(t)/\mu_s(t)$) will be hard to detect because the rate in an epidemic state is similar to the rate in the non-epidemic state. Similarly, more uncertainty in the baseline rate $\mu_s(t)$ may lead to less power in detecting the epidemic disease rate. This section describes a simulation study aimed at determining the sensitivity of the absorbing state model as a surveillance tool under various scenarios.

For this simulation study, assume that an initial training period of $d \in \{1, 3, 7, 14\}$ days is available in which the disease is known to be in a non-epidemic state. To simulate a single data set, an initial d days of data are first simulated under non-epidemic conditions using (2) and (4). After the initial d days, $T = 21$ days are simulated for $S = 100$ spatial locations using the parameter values $(\pi_1, \pi_2) = (0.01, 0.10)$, $\sigma_\theta^2 = \sigma_\xi^2 = 0.05$, $\exp\{\alpha_\mu\} \in \{1, 2, 4, 8\}$, and $\kappa = \exp\{\alpha_\lambda - \alpha_\mu\} \in \{0.2, 0.5, 1.0, 1.5, 2, 4\}$. The parameter κ here represents the expected percentage increase in the epidemic rate from the baseline rate. Twenty-five data sets were simulated for each combination of $(\exp\{\alpha_\mu\}, \kappa, d)$ making, in total, 2400 simulated data sets. The spatial networks for ξ_s and $\theta_s(t)$ were assumed to come from the spatial network of 100 counties with connectivity pattern corresponding to the counties in North Carolina and with weights $w_{\xi,is} = w_{\theta,is}$ defined in (5). No demographic covariates were used for this simulation study. The prior distributions described in Section 3 were used and each data set was fit using MCMC with an initial burn-in period of 2500 draws, after which 10,000 draws were obtained for posterior analysis.

The decision rule used for the simulation study is based on the posterior probability of an epidemic. For example, for prospective analyses, the decision rule is to declare an epidemic at time t if $\Pr(\delta_s(t) = 1 \mid \boldsymbol{Y}(1:t)) > r$ for some r where $\boldsymbol{Y}(1:t)$ is all data up to time t. Similarly,

in a retrospective analysis, an epidemic is declared at time t if $\Pr(\delta_s(t) = 1 \mid \boldsymbol{Y}) > r$ where $\boldsymbol{Y}$ is *all* data. A simple decision rule is to let $r = 0.5$ in which case, if the evidence in favor of an epidemic state is more than the evidence in favor of a non-epidemic state, then an epidemic is declared. However, this rule may not be optimal for all parameter settings. Table 1 displays the value of r which minimizes the misclassification rate of a prospective analysis based on the results from fitting the absorbing state model to the simulated data.

Table 1 shows that higher values of r are preferred for cases where either $\exp\{\alpha_\mu\}$ or κ is low. This result is intuitive in that if $\exp\{\alpha_\mu\}$ or κ are low then the two states are similar and weighty evidence is needed before declaring an epidemic. Based on these findings, the value of r used for the decision rule in the simulation study was determined based on the posterior distribution of model parameters. Specifically, an epidemic was declared if $\Pr(\delta_s(t) = 1 \mid \boldsymbol{Y}) > r$ where

$$r = \begin{cases} 0.75 & \text{if } \widehat{\kappa} < 1, \\ 0.65 & \text{if } 1 \leq \widehat{\kappa} \leq 1.75, \\ 0.50 & \text{otherwise,} \end{cases} \tag{8}$$

and $\widehat{\kappa}$ is the posterior mean of κ.

Figure 1 and 2 display the results from the simulation study based on prospective and retrospective analyses, respectively. Specifically, the misclassification rate (MR) for the prospective analysis in Figure 1 is given by

$$\begin{aligned} \text{MR}_{\text{pro}} = \frac{1}{ST} \sum_{s,t} \mathbb{I}_{\{\Pr(\delta_s(t)=1|\boldsymbol{Y}(1:t))>r\}} \mathbb{I}_{\{\text{No epidemic at location } s \text{ at time } t\}} + \cdots \\ \frac{1}{ST} \sum_{s,t} \mathbb{I}_{\{\Pr(\delta_s(t)=1|\boldsymbol{Y}(1:t))<r\}} \mathbb{I}_{\{\text{Epidemic at location } s \text{ at time } t\}}, \end{aligned}$$

where r is given in (8), and $\boldsymbol{Y}(1{:}t)$ is all data up to time t. Similarly, the misclassification rate for the retrospective analysis in Figure 2 is

$$\begin{aligned} \text{MR}_{\text{ret}} = \frac{1}{ST} \sum_{s,t} \mathbb{I}_{\{\Pr(\delta_s(t)=1|\boldsymbol{Y}(1:T))>r\}} \mathbb{I}_{\{\text{No epidemic at location } s \text{ at time } t\}} + \cdots \\ \frac{1}{ST} \sum_{s,t} \mathbb{I}_{\{\Pr(\delta_s(t)=1|\boldsymbol{Y}(1:T))<r\}} \mathbb{I}_{\{\text{Epidemic at location } s \text{ at time } t\}}. \end{aligned}$$

Holding all else constant, as the global baseline parameter α_μ increases, the empirical misclassification rate decreases in both the prospective and retrospective studies. Similarly, as κ increases, the empirical misclassification rate decreases. Additionally, the rate of decrease in the misclassification increases as both α_μ and κ increase. These findings are reasonable in that if α_μ or κ increases while holding the other constant, the discrepancy between non-epidemic and epidemic states increases because $\alpha_\lambda = \log(\kappa) + \alpha_\mu$.

Somewhat surprisingly, the length of the training period (d) seems to have little effect on misclassification rates. In fact, empirical intervals give evidence that the misclassification rates across each level of d are the same. This result indicates that the model is able to estimate the baseline rate with sufficient accuracy after only a few days of training data (probably because

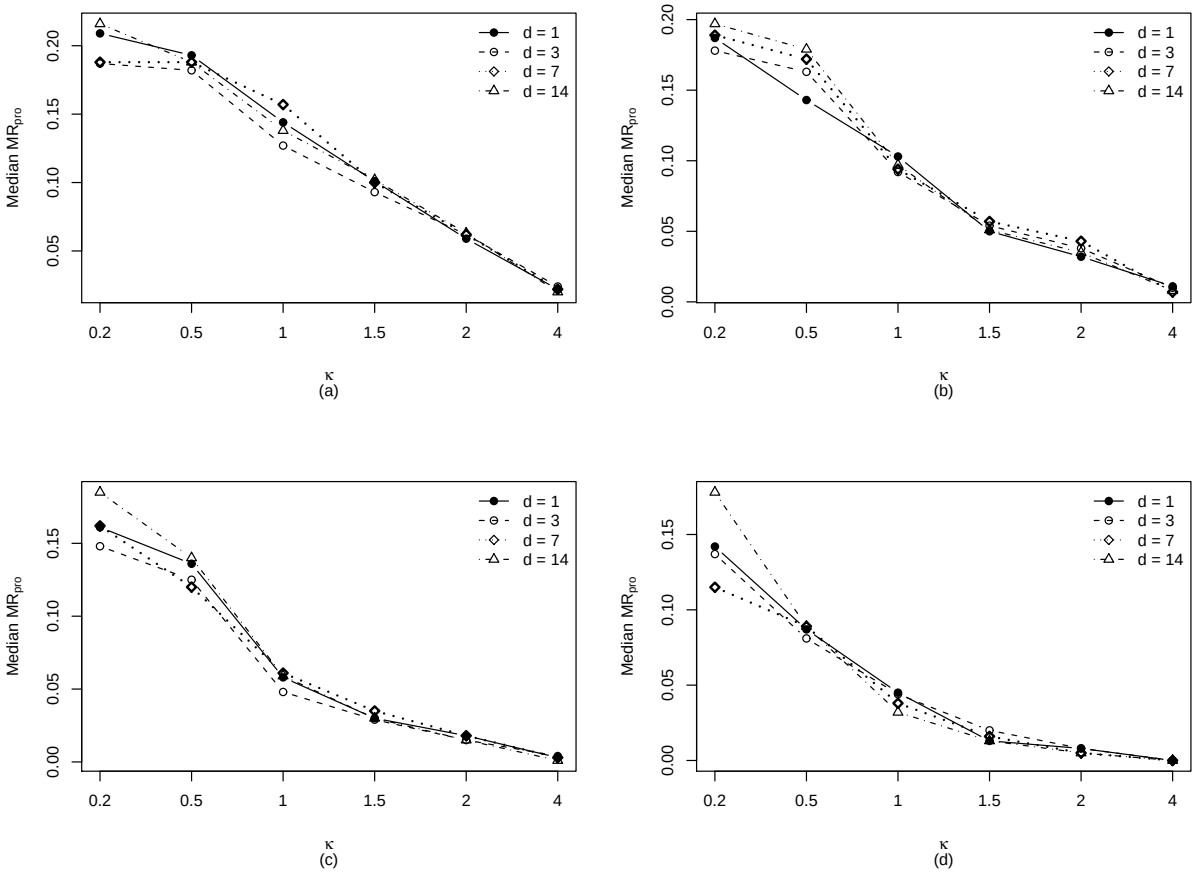

Figure 1. Empirical misclassification rates of the absorbing state model for a prospective analysis when (a) $\alpha_\mu = \log(1)$*, (b)* $\alpha_\mu = \log(2)$*, (c)* $\alpha_\mu = \log(4)$*, and (d)* $\alpha_\mu = \log(8)$*. As* α_μ *and/or* κ *increase, empirical misclassification rates decrease.*

there are a relatively large number of counties, all with empty covariate structure, making it easy to borrow strength across counties in order to estimate the baseline).

One concern is the high misclassification rate when $\kappa = 0.2$. Indeed, when $\kappa = 0.2$ and $\alpha_\mu = \log(1)$, both the prospective and retrospective analyses have misclassification rates near 0.20. However, for cases when κ is small, the problem is nearly equivalent to properly classifying observations as $\mathcal{P}(\exp\{\alpha_\mu\})$ or $\mathcal{P}((1+\kappa)\times\exp\{\alpha_\mu\})$ random variables. For example, when $\alpha_\mu = \log(1)$, then observations are essentially classified as either $\mathcal{P}(1)$ or $\mathcal{P}(1.2)$ random variables, and on information-theoretic grounds, these two distributions are nearly indistinguishable.

5. ILLUSTRATIVE APPLICATION

As an illustration, this section applies the absorbing state model to weekly influenza and pneumonia mortality data from the *Morbidity and Mortality Weekly Report*. The data set is

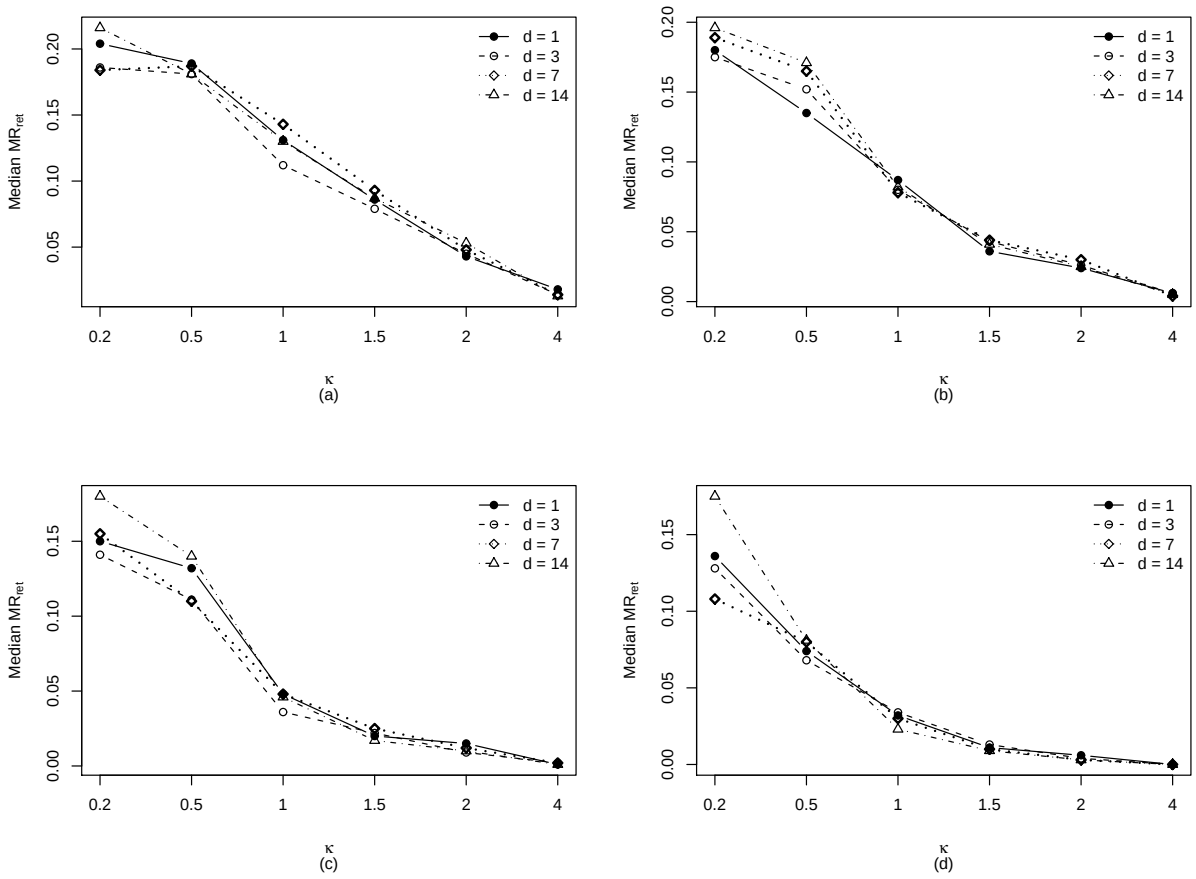

Figure 2. Empirical misclassification rates of the absorbing state model for a retrospective analysis when (a) $\alpha_\mu = \log(1)$*, (b)* $\alpha_\mu = \log(2)$*, (c)* $\alpha_\mu = \log(4)$*, and (d)* $\alpha_\mu = \log(8)$*. As* α_μ *and/or* κ *increase, empirical misclassification rates decrease.*

publicly available at `http://www.cdc.gov/mmwr`. Specifically, the mortality data analyzed here is weekly counts of deaths resulting from influenza or pneumonia in 2009 for 122 cities across the United States; however, due to contiguity reasons, Honolulu, Hawaii was excluded from this analysis. Figure 3 displays the 121 locations considered for this study. Furthermore, because the absorbing state model is only useful for detecting the *first* incidence of an outbreak, the analysis is confined to data collected after the week ending May 2, 2009.

One particular challenge in the mortality data set is that 5% of the data is missing. To complicate matters further, all the data for Fort Worth, TX and New Orleans, LA are missing. Other cities with a noteworthy amount of missing data are Chicago, IL, Detroit, MI, and Baton Rouge, LA with 60%, 29%, and 26% missing data, respectively. Traditional surveillance approaches such as the CUSUM and spatial scan statistic mentioned in the Introduction are not well equipped to deal with missing data. However, from a model-based approach, missing data is treated straightforwardly as an additional unknown. Specifically, with missing data, the distribution of interest

Figure 3. The 121 city locations for the flu and pneumonia mortality study.

is the posterior distribution $[\boldsymbol{Y}_m, \boldsymbol{\Theta} \mid \boldsymbol{Y}_o]$ where $\boldsymbol{Y}_m$ and $\boldsymbol{Y}_o$ represent the missing and observed data, respectively, and $\boldsymbol{\Theta}$ is all model parameters. Using a partially collapsed Gibbs sampler [27] which samples the joint distribution by first obtaining a draw from $[\boldsymbol{\Theta} \mid \boldsymbol{Y}_o]$ and then obtaining a draw from $[\boldsymbol{Y}_m \mid \boldsymbol{\Theta}, \boldsymbol{Y}_o]$ will, typically, be more efficient than sampling the full conditionals $[\boldsymbol{\Theta} \mid \boldsymbol{Y}_m, \boldsymbol{Y}_o]$ and $[\boldsymbol{Y}_m \mid \boldsymbol{\Theta}, \boldsymbol{Y}_o]$. Inference for $\boldsymbol{\Theta}$ is carried out via the marginal posterior distribution $[\boldsymbol{\Theta} \mid \boldsymbol{Y}_o]$ and incorporates the uncertainty associated with the missing data $\boldsymbol{Y}_m$.

Another important aspect of this data set is the availability of covariates. Specifically, population counts for each city can be obtained from census data. To use this covariate information, let $\boldsymbol{x}_s(t) = \boldsymbol{h}_s(t)$ be the (centered) log-population estimate of city s for $s = 1, \ldots, 121$; i.e. $\boldsymbol{x}_s(t) = \boldsymbol{h}_s(t) = \log(\text{pop}_s) - (121)^{-1} \sum_s \log(\text{pop}_s)$ where pop_s is the population of city s.

Despite the availability of some covariate information, the model $\xi_s(t) \equiv \xi_s$ in (4) was used because the spatial region being considered is quite extensive and baseline death rates $(\mu_s(t))$ are likely to vary over the region. The model (6) was again used. As a neighborhood structure, cities s and s' were considered "neighbors" if they are in the same state or in contiguous states. To perform inference, MCMC methodology was used to obtain 25,000 draws from the posterior distribution $[\boldsymbol{Y}_m, \boldsymbol{\Theta} \mid \boldsymbol{Y}_o]$ after discarding an initial 10,000 draws. Convergence diagnostics showed evidence of convergence [28]. For comparison, both a prospective and retrospective analysis were performed.

The mean (95% credible interval) of the posterior distribution for $\kappa = \exp\{\alpha_\lambda - \alpha_\mu\}$ was found to be 0.49 (0.21,0.78) indicating that the outbreak rate was only, on average, 50% higher than the baseline rate. Hence, as discussed in Section 4, higher values of r should be used in the decision rule. Thus, for the prospective study, the cutoff $r = 0.75$ was used; thus the flu season was declared to have started at location s if $\mathbb{P}\text{r}(\delta_s(t) = 1 \mid \boldsymbol{Y}_o(1:t)) > r$. Similarly, for the retrospective analysis, the decision rule was to declare the start of the flu season if $\mathbb{P}\text{r}(\delta_s(t) = 1 \mid$

$\boldsymbol{Y}_o) > r$.

Using the decision rule above, the prospective and retrospective analyses differed on when and where the influenza season started. The prospective analysis first sounded an alarm in Rochester, NY, Des Moines, IA, and Providence, RI, in the week of September 20, 2009. In contrast, the retrospective analysis first sounded an alarm in San Antonio, TX, three weeks earlier (the week of August 30). To investigate this further, for Des Moines, IA, and Providence, RI, averaged 3.33 and 5.67 deaths per week for the three weeks preceding the week of September 20. However, for the week of September 20, Des Moines and Providence saw 10 and 12 deaths from influenza and pneumonia which corresponds to more than a doubling of deaths compared to the previous three weeks. However, in the week following September 20, Des Moines saw 2 deaths and Providence saw 3 suggesting that the high count during the week of September 20 was due to chance or to some non-contagious event. Subsequently, the prospective analysis, after observing the low counts following the week of September 20, "undeclared" Des Moines and Providence to have outbreaks of influenza. This type of behavior for a prospective study using the absorbing state model is encouraging because the model can identify abnormal increases in rate but readjust after more data become available.

To compare the prospective and retrospective analyses, Figures 4 (a), (c), and (e) display the prospective posterior probabilities $[\delta_s(t) = 1 \mid \boldsymbol{Y}_o(1:t)]$ and Figures 4 (b), (d), and (f) display the retrospective posterior probabilities $[\delta_s(t) = 1 \mid \boldsymbol{Y}_o]$ for three selected weeks. Notice during the week of October 4^{th} (Figure 4 (a) and (b)), the prospective and retrospective analyses agree on declaring an outbreak on several locations but the prospective analysis seems to not declare an outbreak for the western cities of San Antonio, TX, Las Vegas, NV, and Sacramento, CA. The prospective analysis didn't declare an outbreak in San Antonio, Las Vegas, and Sacramento until the week of October 7 for all these cities (based on the decision rule above). Notice that for the week of November 15^{th} (Figures 4 (c) and (d)), the prospective analysis declares an influenza outbreak for Spokane, WA but later corrects that decision. Specifically, the prospective analysis declared an influenza outbreak for the entire month of November in Spokane but that decision was later changed when December exhibited lower numbers of deaths from influenza and pneumonia.

To illustrate the usefulness of a model-based approach for dealing with missing data, Figure 5 displays the retrospective reconstruction of the estimated number of deaths from influenza and pneumonia for New Orleans, LA. Specifically, Figure 5 displays the posterior expectation $\mathbb{E}(Y_s(t) \mid \boldsymbol{Y}_o)$ over time for New Orleans. Intuitively, the estimated number of deaths increases as winter approaches indicating that influenza and pneumonia are more prevalent in the winter.

6. DISCUSSION

This article develops the structural framework for disease and syndromic surveillance models. Specifically, an absorbing state model with a novel non-separable space-time neighborhood structure was proposed which focuses on detecting the first incidence of an outbreak. Simulations studies revealed that for very small outbreaks, the epidemic and non-epidemic states were difficult to distinguish. However, for larger epidemics the model performed very well in distinguishing between the two states. Subsequent application to an influenza and pneumonia mortality data set showed that the model was useful for prospective analyses.

The focus of this article was on the surveillance of univariate diseases or syndromes. Mul-

tivariate surveillance models are rarely considered in the literature due to their complexity and difficulties in computation. Additionally, decision rules based on multivariate models can be difficult to construct. However, much power could be gained by jointly modeling several diseases or syndromes, and this might lead to improved methodology with significant practical value.

ACKNOWLEDGMENTS

This research was supported in part by NSF grants DMS–0914906 to the National Institute of Statistical Sciences, DMS–0914903 to Clemson University, DMS–0914603 to the University of Georgia Research Foundation, DMS–0914921 to the University of South Carolina Research Foundation and DMS–0112069 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

APPENDIX: FFBS ALGORITHM

The forward filtering backward sampling (FFBS) algorithm is used to draw $\boldsymbol{\delta}_s = (\delta_s(1), \ldots, \delta_s(T))'$ from its complete conditional distribution. For the absorbing state case, the FFBS algorithm proceeds as follows. Let $[\cdot]$ denote a general probability distribution that is conditional on *all* parameters except $\boldsymbol{\delta}_s$.

1. Calculate
$$\begin{aligned}\Pr(\delta_s(1) = 1 \mid Y_s(1)) &= \frac{[Y_s(1) \mid \delta_s(1) = 1]\gamma_s(0)}{[Y_s(1) \mid \delta_s(1) = 1]\gamma_s(0) + [Y_s(1) \mid \delta_s(1) = 0](1 - \gamma_s(0))} \\ &= p_s(1)\end{aligned}$$
where $\gamma_s(t)$ is given by (7) and $\gamma_s(0) = \pi_1$.

2. For $t = 2, \ldots, T$, let
$$\begin{aligned}\Pr(\delta_s(t) = 1 \mid \boldsymbol{Y}_s(1 : (t-1))) &= \sum_{z=0}^{1} [\delta_s(t) = 1 \mid \delta_s(t-1) = z][\delta_s(t-1) = z \mid \boldsymbol{Y}_s(1 : (t-1))] \\ &= \gamma_s(t-1)(1 - p_s(t-1)) + p_s(t-1) \\ &= d_s(t)\end{aligned}$$
where $\boldsymbol{Y}_s(1 : t) = (Y_s(1), \ldots, Y_s(t))$ and calculate
$$\begin{aligned}\Pr(\delta_s(t) = 1 \mid \boldsymbol{Y}_s(1 : t)) &= \frac{[Y_s(t) \mid \delta_s(t) = 1]d_s(t)}{[Y_s(t) \mid \delta_s(t) = 1]d_s(t) + [Y_s(t) \mid \delta_s(t) = 0](1 - d_s(t))} \\ &= p_s(t)\end{aligned}$$

3. Sample $\delta_s(T)$ from a Bernoulli distribution with parameter $p_s(T)$.

4. For $t = T - 1, \ldots, 1$, draw $\delta_s(t)$ from a Bernoulli distribution with parameter
$$\Pr(\delta_s(t) = 1 \mid \delta_s(t+1), \boldsymbol{Y}_s(1 : t)) = \frac{\mathbb{I}_{\{\delta_s(t+1)=1\}} p_s(t)}{\mathbb{I}_{\{\delta_s(t+1)=1\}} d_s(t+1) + (1 - \mathbb{I}_{\{\delta_s(t+1)=1\}})(1 - d_s(t+1))}.$$

REFERENCES

[1] Page, ES. Continuous inspection schemes. *Biometrika* 1954; **41**:100–115.

[2] Fricker, RD, Hegler, BL, Dunfee, DA. Comparing syndromic surveillance detection methods: Ears versus a cusum-based methology. *Statistics in Medicine* 2008; **27**:3407–3429.

[3] Cowling, BJ, Wong, IOL, Ho, LM, Riley, S, Leung, GM. Methods for monitoring influenza surveillance data. *International Journal of Epidemiology* 2006; **35**:1314–1321.

[4] Rossi, G, Lampugnani, L, Marchi, M. An approximate cusum procedure for surveillance of health events. *Statistics in Medicine* 1999; **18**:2111–2122.

[5] Qiu, P, Hawkins, D. A rank based multivariate cusum procedure. *Technometrics* 2001; **43**:120–132.

[6] Mason, RL, Champ, CW, Tracy, ND, Wierda, SJ, Young, JC. Assessment of multivariate process control techniques. *Journal of Quality Technology* 1997; **29**:140–143.

[7] Kulldorff, M. A spatial scan statistic. *Communications in Statistics - Theory and Methods* 1997; **26**(6):1481–1496.

[8] Walther, G. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics* 2010; **38**(2):1010–1033.

[9] Neill, DB, Cooper, GF. A multivariate bayesian scan statistic for the early event detection of characterization. *Machine Learning* 2010; .

[10] Kulldorff, M, Mostashari, F, Duczmal, L, Yih, WK, Kleinman, K, Platt, R. Multivariate scan statistics for disease surveillance. *Statistics in Medicine* 2007; **26**:1824–1833.

[11] Neill, DB, Moore, AW, Cooper, GF. A bayesian spatial scan statistic. In *Advances in Neural Information and Processing Systems 18*, Weiss, Y, Scholkopf, B, Platt, J, eds. 1003–1010.

[12] LeStrat, Y, Carrat, F. Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in Medicine* 1999; **18**:3463–3478.

[13] Madigan, D. Bayesian data mining for health surveillance. In *Spatial and Syndromic Surveillance for Public Health*, Lawson, AB, Kleinman, K, eds. Wiley, 2005; 203–221.

[14] Martínez-Beneito, MA, Conesa, D, López-Quílez, A, López-Maside, A. Bayesian markov switching models for the early detection of influenza epidemics. *Statistics in Medicine* 2008; **27**:4455–4468.

[15] Rath, TM, Carrerras, M, Sebastiani, P. Automated detection of influenza epidemics with hidden markov models. In *Proceedings of the 5th International Symposium on Intelligent Data Analysis*, Springer-Verlag, ed. 521–531.

[16] Knorr-Held, L, Richardson, S. A hierarchical model for space-time surveillance data on meningococal disease incidence. *Journal of the Royal Statistical Society Series C* 2003; **52**(2):169–183.

[17] Banks, D, Datta, G, Karr, A, Lynch, J, Niemi, J, Vera, F. Bayesian car models for syndromic surveillance on multiple data streams: Theory and practice. *Information Fusion* 2010; **In press**(xx):DOI: 10.1012/j.inffus.2009.10.005.

[18] Held, L, Hofmann, M, Höhle, M. A two-component model for counts of infectious diseases. *Biostatistics* 2006; **7**(3):422–437.

[19] Held, L, Höhle, M, Hofmann, M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 2005; **5**:187–199.

[20] Besag, JE, Green, PJ, Higdon, DM, Mengersen, KL. Bayesian computation and stochastic systems (with discussion). *Statistical Science* 1995; **10**:3–66.

[21] Banerjee, S, Carlin, BP, Gelfand, AE. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2004.

[22] Mugglin, AS, Cressie, N, Gemmell, I. Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine* 2002; **21**:2703–2721.

[23] Besag, J. Spatial interaction and the analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* 1974; **55**:25–37.

[24] Gamerman, D, Lopes, HF. *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 2006.

[25] Haario, H, Saksman, E, Tamminen, J. An adaptive metropolis algorithm. *Bernoulli* 2001; **7**(2):223–242.

[26] Rue, H, Martino, S, Chopin, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B* 2009; **71**(2):319–392.

[27] van Dyk, DA, Park, T. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association* 2008; **103**(482):790–796.

[28] Cowles, MK, Carlin, BP. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* 1996; **91**(434):883–904.

Figure 4. Comparison of prospective and retrospective epidemic alarms for influenza and pneumonia deaths during 2009 according to the absorbing state model. Triangles, diamonds, squares, and circles indicate that the posterior probability of an outbreak is greater than 0.5, 0.7, 0.8, and 0.9, respectively. Small, unfilled circles represent those cities for which the posterior probability of an outbreak is less than 0.5.

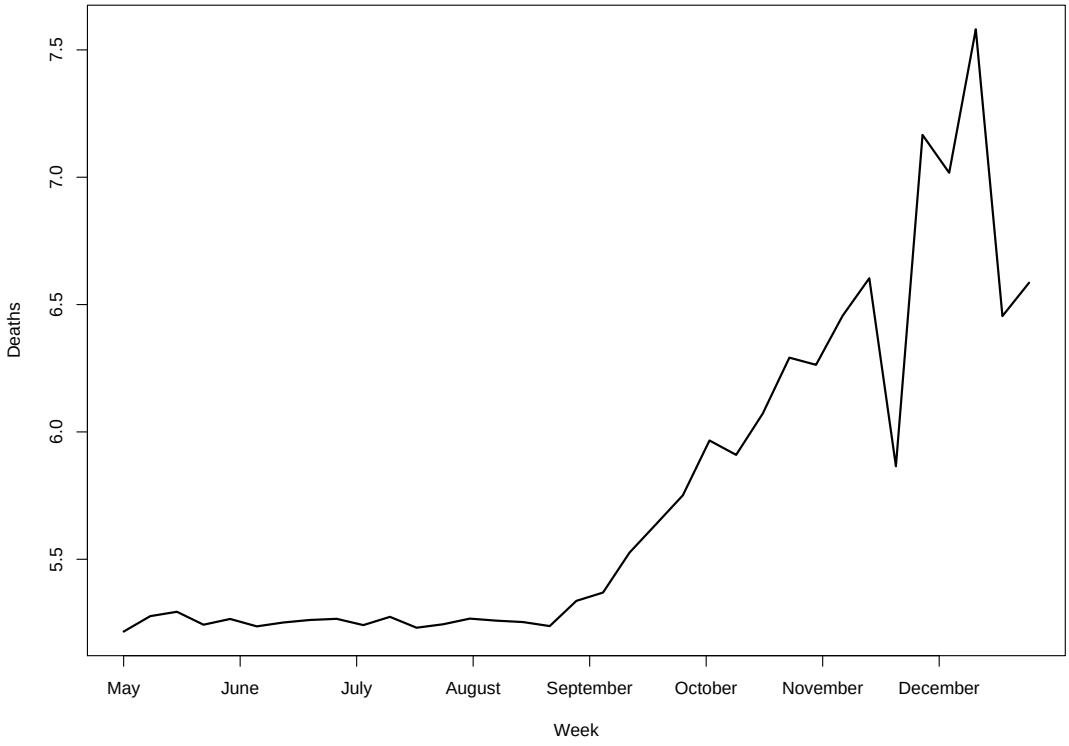

Figure 5. The posterior expectation $\mathbb{E}(Y_s(t) \mid \boldsymbol{Y}_o)$ over time for the city of New Orleans, LA.

Table 1. Values of r which minimize the empirical misclassification rate of the absorbing state model where the decision rule is based on $\mathbb{P}r(\delta_s(t) \mid \boldsymbol{Y}) > r$. For diseases with a small baseline $(\exp\{\alpha_\mu\})$ or small epidemic-to-baseline ratio (κ), larger values of r are preferred in order to declare an epidemic as opposed to larger baselines.

	κ					
$\exp\{\alpha_\mu\}$	0.2	0.5	1	1.5	2	4
1	0.80	0.81	0.80	0.74	0.64	0.50
2	0.80	0.80	0.75	0.61	0.58	0.50
4	0.80	0.75	0.61	0.53	0.53	0.51
8	0.70	0.66	0.53	0.49	0. 51	0.53