2011

Prospective Surveillance of Multivariate Spatial Disease Data

A. Corberan-Vallet
Medical University of South Carolina

A. B. Lawson
Medical University of South Carolina

Paper Number/ Resource Identifier: 11-006

Date: 2011

MUSC Author(s): A. Corberan-Vallet and A.B. Lawson

Paper Title: Prospective Surveillance of Multivariate Spatial Disease Data

Complete Author List: A. Corberan-Vallet and A.B. Lawson

Prospective Surveillance of Multivariate Spatial Disease Data

A. Corberán-Vallet and A.B. Lawson

Division of Biostatistics and Epidemiology

Medical University of South Carolina

135 Cannon St. Suite 303, Charleston, SC 29425

E-mail: corberan@musc.edu, lawsonab@musc.edu

Abstract

Surveillance systems are often focused on more than one disease within a predefined area. On those occasions when outbreaks of disease are likely to be correlated, the use of multivariate surveillance techniques integrating information from multiple diseases allows us to improve the sensitivity and timeliness of outbreak detection. In this paper, we present an extension of the surveillance conditional predictive ordinate to monitor multivariate spatial disease data. The proposed surveillance technique, which is defined for each small area and time period as the conditional predictive distribution of those counts of disease higher than expected given the data observed up to the previous time period, alerts us to both small areas of increased disease incidence and the diseases causing the alarm within each area. We investigate its performance within the framework of Bayesian hierarchical Poisson models using a simulation study. An application to diseases of the respiratory system in South Carolina is finally presented.

Keywords: disease surveillance; multiple diseases; Shared component model; conditional predictive ordinate

1 Introduction

Public health surveillance is defined as the ongoing systematic collection, analysis, and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know[1]. Effective surveillance is then essential to protect public health by rapidly detecting and responding to disease epidemics.

Most work on surveillance methodology has evolved in temporal applications, and so numerous methods including process control charts, temporal scan statistics, time-series methodology, and log-linear and other parametric regression models have been proposed to monitor univariate time series of counts of disease[2]. Because of the growing threat of bioterrorism and an increase in the emergence and reemergence of infectious diseases with pandemic potential, numerous studies have recently been conducted to develop new and improved methods for health surveillance. New statistical methods usually use information on both the time and location of events, and so they offer an improved ability to detect localized events that occur in small regions relative to the surveillance of the total count across a larger region. Testing methods are widely used to detect outbreaks of disease in space and time[3,4]. Recent developments for the analysis of space-time disease surveillance data use a statistical model to describe the behavior of disease over space and time during non-epidemic conditions and the emphasis is placed on detection of unusual departures from predictable patterns based on the estimated model[5−10]. These model-based approaches provide a flexible framework for the inclusion of spatial, temporal, space-time interaction, and possible covariate effects.

Multivariate space-time surveillance data also arise naturally in many public health applications. For instance, disease incidence data are often available by age group, gender and race. On some occasions, a range of different diseases are monitored simultaneously to assess the general health status of a region. Some examples are the monitoring of smoking-related cancers, respiratory diseases or gastrointestinal illnesses. In a syndromic surveillance setting, different syndromes associated with disease are monitored simultaneously to detect outbreaks of disease at the earliest possible time, possibly even before definitive disease diagnoses are obtained. Common syndromes are school and work absenteeism, over-the-counter medication sales, emergency department visits, physician telephone calls, etc. On those occasions, the use of surveillance techniques integrating information from the different data sets is important to achieve higher detection power for events that are present simultaneously in more than one data set. Kulldorff et al.[11] presented an extension of the space-time scan statistic to jointly monitor multiple data sets. The multivariate scan statistic is based on a combined log likelihood which is defined as the sum of the individual log likelihoods for those data sets with more counts than expected in the scanning window. So a signal is generated if a cluster is detected in either one or in a combination of data sets. Further extensions, such as the Bayesian multivariate scan statistic[12], have been proposed since then. Banks et al.[13] have proposed a model-based approach to surveillance of spatial data on multiple diseases. The proposed methodology, which is focused on syndromic surveillance, uses univariate Bayesian hierarchical models to model counts of patients with specific symptoms indicative of the same disease in the absence of an epidemic. Indicator variables modeled as a binary Markov random field are then used to detect the presence of disease in each spatial unit. In that study, the authors assume that an increase in the number of cases is observed for all the symptoms at the same time when the disease is present.

In practice however, the different data sets under study may be influenced by common confounding factors, and so they are likely to be correlated. This suggests that we need to consider multivariate disease models to describe the space-time behavior of diseases. The multivariate conditional autoregressive (MCAR) model[14] and the shared component model[15,16] are the two main approaches to model disease risk correlations across both spatial units and diseases. The main advantage of the shared component model is that it enables estimation of shared and disease-specific spatial patterns.

In this paper a shared component model is used to describe the behavior of diseases under non-epidemic conditions. A novelty of the proposed model formulation is the use of indicator variables, which allow for identification of shared and disease-specific latent spatial fields describing the risk surface for each disease. We show then how the surveillance conditional predictive ordinate (SCPO), which was introduced by Corberán-Vallet and Lawson[17] in a univariate model-based surveillance setting to detect areas of unusual disease aggregation, can be straightforwardly extended to incorporate information from multiple diseases. In particular, we define the multivariate surveillance conditional predictive ordinate (MSCPO) for each small area and time period as the conditional predictive distribution of those counts higher than expected given the data collected so far. A parallel surveillance approach across the different areas under surveillance is then carried out, where in each area alarms are sounded if the corresponding MSCPO value is below a specified critical value. This surveillance technique alerts us to both spatial units of increased disease incidence in need of further investigation and the diseases causing the alarm within each area, and consequently it facilitates a timely and informed public health response.

This paper is organized as follows. In Section 2, we present our modeling framework. In Section 3, we review the surveillance conditional predictive ordinate and introduce its multivariate extension to multiple disease surveillance. Section 4 shows the results obtained in a simulation study. The surveillance technique is then applied to emergency room discharges for diseases of the respiratory system in South Carolina. Finally, we conclude with a general discussion of the proposed technique and provide directions for future research.

2 Modeling of endemic periods

2.1 The convolution model

Let y_{it} and e_{it} denote, respectively, the observed and expected count of disease in area i and time period t, for $i = 1, 2, \ldots, m$ and $t = 1, 2, \ldots, T$. We assume here that the observed counts are Poisson distributed

$$y_{it} \sim Po(e_{it}\theta_{it})$$

where θ_{it}, which is often termed the relative risk, represents the excess risk within area i at time t. This component is usually the focus of interest, and so a wide range of spatiotemporal models have been developed to estimate the true relative risk of a disease of interest across a geographic study region. The most common approach to relative risk modeling is to assume a logarithm link to a linear predictor which is a function of fixed observed covariates and spatial, temporal and space-time interaction random effects[18,19].

In a surveillance context, however, the emphasis is placed on detection of changes. To this end, Lawson[20] emphasized the need for a relatively simple model capturing the normal historical variation in disease incidence without absorbing changes in the model fit. In a recent study, Corberán-Vallet and Lawson[17] have demonstrated that the use of a spatial-only model where the relative risks are assumed to be constant over time may improve outbreak detection capability. Hence, we assume that under non-epidemic conditions $\theta_{it} = \theta_i$ for all t, and so unusual departures from predictable patterns based on the overall spatial risk surface are attributable to epidemic processes. To capture spatial correlation in disease maps, we use the convolution model originally proposed by Besag et al.[21]. This model, denoted here by BYM model, assumes that the logarithm of the relative risk is decomposed as

$$log(\theta_i) = \rho + u_i + v_i \tag{1}$$

where ρ is the overall level of the relative risk in the study region, and u_i and v_i represent, respectively, spatially correlated and uncorrelated random effects. As a prior distribution for the intercept we assume a conventional zero-mean Gaussian distribution with variance σ^2_ρ. We use an improper conditional autoregressive (CAR) model[21] as a prior distribution for the correlated heterogeneity, that is

$$u_i | u_{(i)} \sim N\left(\frac{1}{m_i}\sum_{j\in n_i} u_j, \frac{\sigma^2_u}{m_i}\right)$$

where $u_{(i)} = (u_1, u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_m)'$, n_i is the set of spatial neighbors of the ith region, m_i is the cardinality of n_i, and σ^2_u is the correlated spatial component variance. Here the neighborhood is assumed to consist of spatially adjacent areas, but more general definitions (using, for instance, intercentroidal distances) are also possible. The prior distribution for the uncorrelated heterogeneity is the zero-mean Gaussian distribution with variance σ^2_v

$$v_i \sim N(0, \sigma^2_v).$$

2.2 The shared component model

In public health it is often appropriate to consider the analysis of spatially aggregated data on multiple diseases. On those occasions, the use of multivariate models accounting for correlations across both diseases and locations may provide a better description of the data and enhance comprehension of disease dynamics. Knorr-Held and Best[15] introduced a shared component model for the joint spatial analysis of two related diseases where the underlying risk surface for each disease is separated into a shared and a disease-specific component.

These components can be interpreted as surrogates for spatially structured unobserved covariates that are either shared by both diseases or specific to one of the diseases. For the joint analysis of more than two diseases, Held et al.[16] proposed a generalized shared component model where latent spatial fields may be shared by some of the diseases or may enter only in one of the diseases. Assume that there are K diseases and a fixed study region common to all the diseases. Let y_{ik} and e_{ik} be the observed and expected count of disease during a fixed temporal period and θ_{ik} the relative risk, where $i = 1, 2, \ldots, m$ represents the areal unit and $k = 1, 2, \ldots, K$ the disease. The extended shared component model is defined as

$$\begin{aligned} & y_{ik} \sim Po(e_{ik}\theta_{ik}) \\ & log(\theta_{ik}) = \rho_k + \sum_j \delta_{j,k} w_{j,i} \end{aligned} \tag{2}$$

where $w_j = (w_{j,1}, w_{j,2}, \ldots, w_{j,m})'$ denotes the jth spatial random effect, and the scaling parameter $\delta_{j,k}$ determines the relative contribution of the spatial random effect to disease k. For each spatial field w_j, it is assumed that the terms $log(\delta_{j,1}), log(\delta_{j,2}), \ldots, log(\delta_{j,n_{w_j}})$ follow a multivariate Gaussian distribution with mean zero and marginal variance $\sigma^2_{\delta_j}$, but under the restriction that

$$\sum_{l=1}^{n_{w_j}} \log(\delta_{j,l}) = 0$$

n_{w_j} being the number of relevant diseases for w_j. Consequently, this model formulation requires the prespecification of the number of spatial random effects and the diseases relevant for each one of them. In practice, however, this will not always be known in advance. The number of possible shared and disease-specific components increases rapidly with the number of diseases under study, and so numerous model formulations become possible. MacNab[22] emphasized the need for a careful and realistic formulation of common risk factors. Because dependencies between disease risks are given a priori in Model (2), an inappropriate formulation of shared and disease-specific components can lead to misspecification of the latent spatial fields, lack of model identifiability and failure of MCMC convergence.

Different variants of the above shared component model have been used to model correlations both between and within areal units. For instance, Ma and Carlin[23] replace (2) with

$$log(\theta_{ik}) = \delta_k w_i + \psi_{ik} \tag{3}$$

where the term ρ_k is not included in the model because the expected counts are age-adjusted internally. Similar to the generalized common spatial factor model introduced by Wang and Wall[24], a single spatial random effect is used to model the correlation between diseases and locations. The scaling parameters δ_k allow different risk gradients for different diseases. To avoid identifiability problems, δ_K is set equal to 1, while the remaining scaling parameters are assumed to be unconstrained. The residuals ψ_{ik} are originally assumed to be independent across both areas and diseases, that is $\psi_{ik} \sim N(0, \sigma^2_{\psi_k})$, although they can be generalized to independent CAR models. A similar model is used in Oleson et al.[25] to investigate the spatial and temporal variation in lung, oral and esophageal cancer rates in Iowa. In that study a latent temporal process is incorporated into Model (3) to allow for temporal variation.

In our surveillance setting, disease maps which have an associated temporal dimension are analyzed prospectively with the objective of detecting changes in the risk pattern of diseases. Hence, for each area i and time period t there is a vector of K counts of disease. As in the univariate case, we assume constant relative risks during non-epidemic periods, and so at the first level of the hierarchy counts of disease have a Poisson distribution with mean $e_{itk}\theta_{ik}$. At the second level of the hierarchy the log relative risks are modeled as

$$log(\theta_{ik}) = \rho_k + \sum_{l=1}^{L} \phi_{l,k}\, \delta_{l,k}\, w_{l,i} + \psi_{ik} \tag{4}$$

where ρ_k is the disease-specific overall risk; L represents the number of spatial fields $w_l = (w_{l,1}, w_{l,2}, \ldots, w_{l,m})'$ needed to describe the correlation across both areas and diseases; $\phi_{l,k}$ is a binary indicator variable that takes the value one if the spatial random effect w_l has an influence on disease k and the value zero otherwise; $\delta_{l,k}$ is the scaling parameter that measures the contribution of w_l to disease k, and ψ_{ik} is the uncorrelated term, which is assumed to be zero-mean Gaussian distributed, $\psi_{ik} \sim N(0, \sigma^2_{\psi_k})$.

In general, the number of components (L) is not known, and so it must be estimated. There are several different procedures to the estimation of L. A simple approach, which has been successfully implemented in related studies, is to assume a large number of L components a priori. The presence of each latent component is then determined based on the posterior mean of the associated indicator variables[26,27]. As a prior distribution for $\phi_{l,k}$, we consider the Bernoulli distribution with probability p_l, which can be assumed to be constant or can have a hyperprior distribution, for instance the Beta distribution.

The latent spatial fields are assumed to be independent, with each following a CAR prior distribution, that is

$$w_{l,i} | w_{l,(i)} \sim N\left(\frac{1}{m_i} \sum_{j \in n_i} w_{l,j}, \frac{\sigma^2_{w_l}}{m_i} \right).$$

In order to avoid identifiability problems, we set $\sigma^2_{w_l} = 1$, for $l = 1, 2, \ldots, L$, so that the variance of $\delta_{l,k}\, w_{l,i}$ is determined by $\delta_{l,k}$[28]. As a prior distribution for the scaling parameters $\delta_{l,k}$, which can then be assumed to be unconstrained, we use a non-informative zero-mean Gaussian distribution.

Similar to the model proposed by Held et al.[16], the proposed shared component model assumes that there may be more than one latent spatial field which can be shared by some of the diseases or may be relevant only to one of them. However, by using indicator variables in the model formulation, it is not necessary to specify the structure of the multivariate model in advance.

3 Detection of epidemics: The multivariate surveillance conditional predictive ordinate

The conditional predictive ordinate (CPO) was first defined by Geisser[29] as the posterior predictive distribution of the observation y_i when the model is fitted to all data except y_i. That is,

$$\text{CPO}_i = f(y_i|y_{(i)}) = \int f(y_i|\varphi, y_{(i)})\pi(\varphi|y_{(i)})d\varphi$$

where $y_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ is the data vector with y_i deleted. Small CPO values, which indicate a poor fit by the model, can be used to detect observations discrepant from the given model. The CPO has been widely used in the statistical literature as a Bayesian model assessment tool in different contexts[30]. Recently, Corberán-Vallet and Lawson[17] adapted the CPO in a surveillance context to detect small areas of unusual disease incidence. Let $y_t = (y_{1t}, y_{2t}, \ldots, y_{mt})'$ be the vector of disease counts observed at time period t, $y_{1:t-1} = (y'_1, y'_2, \ldots, y'_{t-1})'$ the vector of all the data observed up to time $t-1$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$ the relative risk vector under non-epidemic conditions. The surveillance CPO (SCPO) is defined for each small area i and time period t as

$$\begin{aligned} \text{SCPO}_{it} &= f(y_{it}|y_{1:t-1}) = \int f(y_{it}|\theta_i, y_{1:t-1})\pi(\theta_i|y_{1:t-1})\, d\theta_i \\ &\approx \frac{1}{J}\sum_{j=1}^{J} Po(y_{it}|e_{it}\theta_i^{(j)}) \end{aligned} \quad (5)$$

where $\{\theta_i^{(j)}\}_{j=1}^{J}$ is a set of relative risks sampled from the posterior distribution that corresponds to the previous time period. The main difference with respect to the CPO is that the SCPO is calculated using only data from previous time points. This is fundamental in a surveillance context, since the inclusion of observations from the new time period may lead to a different model for the relative risk pattern. Hence, if no change in risk takes place at time t, the relative risk in area i and time t, θ_{it}, is equal to θ_i and the observation y_{it} is representative of the data expected under the previously fitted model. Otherwise, SCPO values close to zero are obtained.

In order to detect as early as possible emerging outbreaks of disease, SCPO values are calculated each time new observations become available. An alarm is then generated for the ith small area at time t if the corresponding SCPO value is below a specified critical value α and $y_{it} > e_{it}\hat{\theta}_i$, $\hat{\theta}_i$ being the posterior mean of the relative risk at the previous time period. Since the value of the

SCPO depends on the mean of the Poisson distribution, it is necessary to scale the SCPO to use the same critical value for all the areas. For the CPO, Congdon[31] recommends scaling the CPO values by dividing by their maximum and considering as outliers those observations with a low CPO value, for instance below 0.01. In the surveillance setting, a scaled SCPO can be defined as[17]

$$\text{sSCPO}_{it} = \frac{\text{SCPO}_{it}}{f(e_{it}\hat{\theta}_i | y_{1:t-1})},$$

so that it takes values close to one if the observation at time t is close to the data expected under the previously fitted model, and values close to zero otherwise.

In the multivariate surveillance setting, spatial data on multiple diseases are observed at each time period, and a decision concerning whether a disease incidence has increased has to be made sequentially based on the data collected so far. We believe that a global increase in the incidence of a disease in all the areas occurring at the same time point is unlikely. Similarly, disease outbreaks need not necessarily occur at the same time for all the diseases under surveillance or affect the same spatial units. So, for each area i and time t, let $y_{it} = (y_{it1}, y_{it2}, \ldots, y_{itK})$ be the vector of observed counts of disease, $e_{it} = (e_{it1}, e_{it2}, \ldots, e_{itK})$ the vector of expected counts, $\hat{\theta}_i = (\hat{\theta}_{i1}, \hat{\theta}_{i2}, \ldots, \hat{\theta}_{iK})$ the vector of posterior relative risk estimates at the previous time point, and $y_{it}^h = (y_{itk_1}, y_{itk_2}, \ldots, y_{itk_n})$ the vector of observed counts higher than expected, that is $y_{itk} > e_{itk}\hat{\theta}_{ik}$. A multivariate extension of the SCPO incorporating information from multiple diseases can be defined as

$$\begin{aligned} \text{MSCPO}_{it} &= f(y_{itk_1}, y_{itk_2}, \ldots, y_{itk_n} | y_{1:t-1}) \\ &= \int\int \ldots \int f(y_{itk_1}, y_{itk_2}, \ldots, y_{itk_n} | \theta_{ik_1}, \theta_{ik_2}, \ldots, \theta_{ik_n}, y_{1:t-1}) \times \\ &\qquad \pi(\theta_{ik_1}, \theta_{ik_2}, \ldots, \theta_{ik_n} | y_{1:t-1}) d\theta_{ik_1}\, d\theta_{ik_2} \ldots d\theta_{ik_n} \end{aligned} \quad (6)$$

if y_{it}^h is not null, and MSCPO_{it} equal to one otherwise. Values of the MSCPO close to zero indicate then unusually high disease counts. Note that when $y_{it}^h = \{y_{itk_1}\}$, the MSCPO_{it} corresponds to the SCPO_{it} for disease k_1. When $n \geq 2$, counts of disease higher than expected are looked at in conjunction to improve the outbreak detection capability.

The multiple integral in (6) does not have a closed form solution, and so simulation is required. A Monte-Carlo approximation to the MSCPO_{it} can be obtained from a posterior sampling algorithm as

$$\frac{1}{J}\sum_{j=1}^{J} Po(y_{itk_1} | e_{itk_1}\theta_{ik_1}^{(j)}) \times Po(y_{itk_2} | e_{itk_2}\theta_{ik_2}^{(j)}) \times \ldots \times Po(y_{itk_n} | e_{itk_n}\theta_{ik_n}^{(j)}) \quad (7)$$

where $\{(\theta_{ik_1}^{(j)}, \theta_{ik_2}^{(j)}, \ldots, \theta_{ik_n}^{(j)})\}_{j=1}^{J}$ is a set of relative risks sampled from the posterior distribution at time $t-1$.

As in the univariate surveillance setting, effective measures based on the MSCPO values have to be constructed to assess if there is any outbreak of disease occurring at time period t. To make MSCPO values comparable across areas and time periods, we propose to consider the scaled MSCPO given by

$$\text{sMSCPO}_{it} = \frac{\text{MSCPO}_{it}}{f(e_{itk_1}\hat{\theta}_{ik_1}, e_{itk_2}\hat{\theta}_{ik_2}, \ldots, e_{itk_n}\hat{\theta}_{ik_n}|y_{1:t-1})} \tag{8}$$

and to perform a parallel surveillance approach across the different areas under surveillance, where an alarm is sounded for the ith small area at time t if the corresponding sMSCPO_{it} is below a specified critical level α. It is important to emphasize here that the proposed surveillance technique alerts us to both small areas of increased disease incidence in need of further investigation and the diseases causing the alarm within each area.

The surveillance technique described herein can be run until the first outbreak is detected and medical intervention takes place. However, it may be of interest to continue the monitoring process to detect either further changes in disease incidences or the end of an epidemic. Corberán-Vallet and Lawson[17] show how the first goal can be achieved by sequentially estimating the model describing the normal behavior of disease using only the last observations. This procedure allows the spatial effects to adapt quickly to changes in the relative risk pattern of a disease, and so it facilitates detection of additional changes in the disease incidence. The second goal may be more relevant in the monitoring of infectious diseases. In order to detect the end of an epidemic, the model describing the behavior of disease in space and time have to be estimated using only counts of disease corresponding to non-epidemic conditions. This can be achieved by assuming that observations detected as unusual are missing when they become part of the history. MSCPO values close to one after consecutive values close to zero are then indicative of the end of an epidemic.

4 Simulation study

In this section, we present a simulation study to assess the performance of the proposed surveillance technique for outbreak detection. The development of a realistic simulation study is important. Here we used the US state of California, which consists of $m = 58$ counties, as the base map to generate counts of diseases at county level for $T = 20$ time periods and $K = 3$ diseases. The total number of viral meningitis cases in California in 2010, which is available from the California department of public health (http://www.cdph.ca.gov), was used to calculate monthly expected counts for the mapped area and Disease 1. Viral meningitis is a relatively common but rarely serious infection of the fluid in the spinal cord and the fluid that surrounds the brain. There is no specific treatment for viral meningitis, which is usually mild and clears up in about a week. It often remains undiagnosed because its symptoms can be similar to those of the common flu: fever, headache, stiff neck, and tiredness. A total of 2623 cases of viral meningitis were diagnosed in California in 2010. Disease rates for the other two diseases were simulated as

$$\begin{aligned} r_2 &= r_1 + Ga(3,1) \\ r_3 &= r_1 + Ga(1,1) \end{aligned}$$

where $r_1 = 0.5656$ is the monthly viral meningitis rate.

True relative risks under non-epidemic conditions were simulated using two different relative risk models. In Scenario 1 we assumed that the three diseases shared a common spatial field, while independent diseases were assumed in Scenario 2. Outbreaks of disease of different intensities were then generated using the expected counts of disease and the simulated relative risks as detailed below.

Scenario 1:

$$log(\theta_{itk}) = \rho_k + w_i + w_{k,i} + \psi_{ik} + \delta_{itk} \tag{9}$$

where $i = 1, 2, \ldots, 58$ denotes the county, $t = 1, 2, \ldots, 20$ the time, and $k = 1, 2, 3$ the disease; $\rho_k \sim N(0, \sigma^2_{\rho_k})$ is the disease-specific overall risk; The components $w = (w_1, w_2, \ldots, w_m)'$ and $w_k = (w_{k,1}, w_{k,2}, \ldots, w_{k,m})'$ represent spatially correlated random effects, each one of them following a CAR model with variance σ^2_w and $\sigma^2_{w_k}$, respectively; $(\psi_{1k}, \psi_{2k}, \ldots, \psi_{mk})'$ is assumed to be a realization of a multivariate Gaussian distribution with zero mean vector and covariance matrix $\sigma^2_{\psi_k} I_m$, and each $\delta_{ik} = (\delta_{i1k}, \delta_{i2k}, \ldots, \delta_{iTk})'$ is assumed to follow a random walk independently of all other counties and diseases, that is $\delta_{itk} \sim N(\delta_{i,t-1,k}, \sigma^2_{\delta_k})$. The values of the standard deviances were $(\sigma_{\rho_1}, \sigma_{\rho_2}, \sigma_{\rho_3}) = (0.01, 0.02, 0.01)$, $(\sigma_w, \sigma_{w_1}, \sigma_{w_2}, \sigma_{w_3}) = (0.1, 0.02, 0.05, 0.05)$, $(\sigma_{\psi_1}, \sigma_{\psi_2}, \sigma_{\psi_3}) = (0.2, 0.1, 0.15)$, and $(\sigma_{\delta_1}, \sigma_{\delta_2}, \sigma_{\delta_3}) = (0.01, 0.02, 0.025)$. Note that the simulated disease risks under non-epidemic conditions were allowed to vary over both space and time.

At time $t_0 = 15$, an epidemic was assumed to start in Los Angeles county ($i_0 = 19$) for the three diseases. Initial expected increases in disease counts due to the epidemic were simulated as

$$I_{i_0 t_0 k} = c_k \, e_{i_0 t_0 k} \, \theta_{i_0 t_0 k} \tag{10}$$

where $c_1 = 0.3$, $c_2 = 0.1$, and $c_3 = 0.5$; that is, at time $t_0 = 15$, a percentage increase in the mean of the Poisson distribution equal to 0.3 was simulated for Disease 1 and so on. At time $t_1 = 17$ the epidemic was assumed to spread to seven neighboring counties (see Figure 1). Increases in disease counts at time t_1 for the affected counties were generated as those in (10)

$$I_{it_1k} = \begin{cases} c_k \, e_{it_1k} \, \theta_{it_1k} & \text{if } i \in R_1 = \{15, 30, 33, 36, 37, 42, 56\} \\ 0 & \text{otherwise} \end{cases}$$

Expected increases at subsequent time periods were assumed to be proportional to those observed at the previous time point, that is $I_{itk} = \beta_{ik} I_{i,t-1,k}$, for $t = 18, 19, 20$. For simplicity, we assumed here $\beta_{i1} = 1.2$, $\beta_{i2} = 1.1$, and

$\beta_{i3} = 1.2$ for all $i \in \{19, R_1\}$. Simulation of outbreaks using realistic models of epidemic progressions is important. However, to assess the capability of surveillance techniques to outbreak detection, it is usually sufficient to use a simple linear model with a moderate slope.

Scenario 2:

$$\begin{aligned}
log(\theta_{it1}) &= \rho_1 + \psi_{i1} + \delta_{it1} + I_{A_1}(i)\, log(1.3) + I_{A_2}(i)\, log(1.8) \\
log(\theta_{it2}) &= \rho_2 + \psi_{i2} + \delta_{it2} + I_{A_3}(i)\, log(1.5) \\
log(\theta_{it3}) &= \rho_3 + \psi_{i3} + \delta_{it3} \qquad (11)
\end{aligned}$$

where parameters ρ_k, ψ_{ik} and δ_{itk} were defined as those in (9). Spatial correlation in model (11) is introduced by three disjoint sets of neighboring counties of higher risk.

At time $t_0 = 15$, an outbreak was generated for eight counties and Diseases 1 and 3 (see Figure 1). Expected increases in disease counts were simulated as

$$\begin{aligned}
I_{it_0k} &= c_k\, e_{it_0k}\, \theta_{it_0k} \\
I_{itk} &= \beta_{ik} I_{i,t-1,k}
\end{aligned}$$

for $i \in R_2 = \{15, 19, 30, 33, 36, 37, 42, 56\}$, $t = 16, 17, \ldots, 20$, and $k = 1, 3$. We assumed $(c_1, c_3) = (0.2, 0.5)$ and $(\beta_{i1}, \beta_{i3}) = (1.2, 1.3)$. At time $t_1 = 17$, an outbreak of Disease 2 was simulated in 5 different counties ($R_3 = \{1, 41, 43, 44, 50\}$). A percentage increase in the mean of the Poisson distribution of 0.3 was assumed initially. Subsequent increases were defined as $I_{it3} = 1.25 I_{i,t-1,3}$, for $i \in R_3$ and $t = 18, 19, 20$.

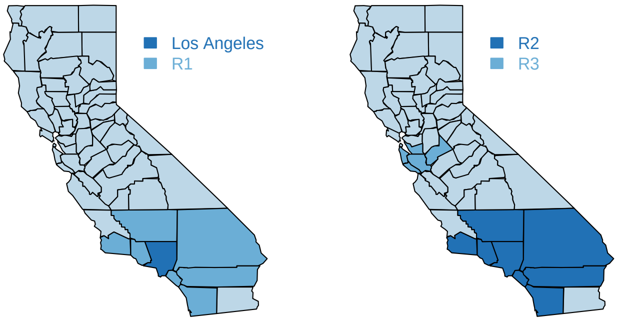

Figure 1: Simulation study. Regions where outbreaks of disease were simulated. Left: Scenario 1. Right: Scenario 2.

Once the values for the expected counts, relative risks, and expected increases in disease counts due to epidemics were specified, we generated the observed counts in the mapped area using the Poisson distribution

$$y_{itk} \sim Po(e_{itk}\theta_{itk} + I_{itk})$$

where $I_{itk} = 0$ for those counties and time periods that do not correspond to epidemics waves. To allow for sampling variability, we simulated 300 data set for each scenario.

The first step in the analysis of the data is to select the model describing the behavior of diseases under non-epidemic conditions. Simulated data under both scenarios were fitted to Model (4) with $L = 6$ latent spatial fields. Posterior sampling was carried out using MCMC with an initial burn-in period of 50000 iterations to assess the convergence of MCMC chains. One posterior sample in five iterations was kept after the burn-in period until a set of 5000 iterations was obtained. A range of different hyperprior specifications of parameter p_l were experimented with. We found that priors penalizing larger values of the number of latent spatial fields, such as the $Be(a, l)$ or the $Exp(a\,l)$ for parameter $p_l/(1 - p_l)$, provide more satisfactory results in general. The results presented here correspond to the case where a $Be(1, l)$ prior is used for parameter p_l, so that as l increases the distribution of p_l gets more concentrated around its mean, which in turn tends to zero. This choice is a compromise between allowing for disaggregation of the underlying risk surface for each disease into different latent spatial fields and searching for a parsimonious fit. Following Ma and Carlin[23], $N(0, 100)$ priors were assumed for the scaling parameters. As a prior distribution for the unknown precision parameters, we used the $Ga(2, 0.5)$, which provides reasonable non-informativeness. Here we accept a latent component in the model if there is at least one associated indicator variable larger than 0.5.

In Scenario 1 a large part of the variation in the data comes from disease-specific components, specifically from the uncorrelated terms. This complicates the detection and proper estimation of the shared latent component. Nevertheless, the selected model generally includes four spatial components, one that is shared by the diseases and three disease-specific CAR components. Table 1 shows the mean square error (MSE) of the relative risks estimates obtained, for each disease, with the shared component model and the overall DIC, averaged over the 300 data sets. For comparative purposes, we also include those results obtained when the diseases are modeled separately by using the convolution model. As can be seen, when the diseases of interest share common risk factors, the use of the shared component model provides more accurate risk estimates and a better fit. In Scenario 2, three disease-specific spatial components were selected.

	Disease 1	Disease 2	Disease 3	DIC
Shared component model	0.035	0.013	0.029	6068.29
Convolution model	0.042	0.018	0.034	6084.30

Table 1: Simulation study. Mean square error of the relative risk estimates obtained, for each disease, with both the shared component model and the convolution model and the overall DIC. The results are averaged over the 300 data sets simulated under Scenario 1.

We next show the results obtained in the prospective analyses of the simulated data with the proposed surveillance technique. Based on the previous results, we used the shared component model with four spatial components to describe the behavior of diseases under non-epidemic conditions in Scenario 1. In Scenario 2 separate convolution models were sequentially fitted to model disease incidences. The relative risks estimates obtained at each time point with the corresponding model were used to calculate the MSCPO values for the new data. Because we are interested in detecting all the areas of increased disease incidence at each time period, we consider the sensitivity, specificity and median time to outbreak detection (MTD) as measures of performance. The sensitivity is defined as the proportion of all the areas undergoing a change in risk that signal an alarm at any time during the outbreak period. The specificity is given by the proportion of in-control areas that are correctly identified as such, that is

$$\begin{aligned} \text{Sensitivity} &= \frac{TA}{TA + FNA} \\ \text{Specificity} &= \frac{TNA}{FA + TNA} \end{aligned}$$

where TA, FA, TNA, and FNA represent, respectively, true alarms, false alarms, true no alarms, and false no alarms during the outbreak period. Finally, let us define, for each small area undergoing an outbreak, the time to outbreak detection as the number of time periods from the beginning of the outbreak until the first alarm is sounded. An infinite time to detection is assigned if no alarm is sounded. The MTD is then defined as the median of the times to detection of those areas of increased disease incidence. It is worthy to emphasize here that a MTD equal to infinite does not mean that no alarm has been sounded, but that the surveillance technique has not detected at least half of the areas of increased disease incidence. The decision rule used in this simulation study was to signal an alarm for the ith area at time t if the $\text{sMSCPO}_{it} < 0.5 \times 10^{-n_{it}}$, n_{it} being the number of counts higher than expected in area i and time t. So, if there is only one count of disease higher than expected in area i and time t the critical value is equal to 0.05; when two counts of disease are higher than expected the critical value is 0.005, and so on. These values were chosen to assure a specificity around 95% for all the diseases and scenarios. Tables 2 and 3 show the sensitivity and MTD of the proposed surveillance technique. Note that one measure value is obtained for each data set. The results presented here are averaged over the 300 data sets simulated for each scenario. For comparative purposes, we also include the results obtained when the diseases were monitored separately by using the SCPO. In this case, an alarm was sounded for area i at time t if the $\text{SCPO}_{it} < 0.05$.

As expected, the SCPO achieves timely detection when changes in disease risks are substantial enough. For Disease 3, an initial percentage increase in the mean of the Poisson distribution equal to 0.5 was simulated in both scenarios. In this case, the outbreak detection capability of both the SCPO and MSCPO is similar. Both surveillance techniques provide also similar results when an outbreak is present in only one disease. This is the case of Disease 2 in Scenario

	SCPO		MSCPO	
	Sens	MTD	Sens	MTD
Disease 1	0.46	Inf	0.68	2
	[0.13,0.75]	[1,Inf)	[0.38,0.94]	[0,Inf)
Disease 2	0.28	Inf	0.60	2.5
	[0.13,0.5]	(Inf,Inf)	[0.25,0.88]	[0,Inf)
Disease 3	0.78	1	0.82	1
	[0.5,1]	[0,Inf)	[0.5,1]	[0,Inf)

Table 2: Simulation study, Scenario 1: Sensitivity and median time to outbreak detection (both posterior average estimates and 95% credible intervals) of the surveillance conditional predictive ordinate (SCPO) and the multivariate surveillance conditional predictive ordinate (MSCPO).

	SCPO		MSCPO	
	Sens	MTD	Sens	MTD
Disease 1	0.42	Inf	0.87	2.5
	[0.13,0.63]	[3,Inf)	[0.63,1]	[1,4]
Disease 2	0.67	2	0.67	2
	[0.4,1]	[0,Inf)	[0.4,1]	[0,Inf)
Disease 3	0.96	1	0.97	1
	[0.81,1]	[0,2]	[0.88,1]	[0,2]

Table 3: Simulation study, Scenario 2: Sensitivity and median time to outbreak detection (both posterior average estimates and 95% credible intervals) of the surveillance conditional predictive ordinate (SCPO) and the multivariate surveillance conditional predictive ordinate (MSCPO).

2. However, by integrating information from multiple diseases, the MSCPO improves considerably the sensitivity and timeliness of event detection when outbreaks of disease occur simultaneously in more than one disease and the proportional increase in disease counts during the epidemic stage relative to the non-epidemic stage is small. For instance, a percentage increase in the mean of the Poisson distribution equal to 0.1 was simulated for Disease 2 in Scenario 1. Counts of disease before and at the onset of the epidemic were then simulated, respectively, from the $Po(e_{it2}\,\theta_{it2})$ and $Po(e_{it2}\,\theta_{it2}\,(1+0.1))$ distributions, which are not different enough to cause an alert when the disease is monitored separately. Hence, only 28% of the areas undergoing an outbreak are detected based on the SCPO. However, the MSCPO signals an alarm for 60% of those areas of increased disease incidence and reduces the MTD to 2.5 units.

5 Case study

This section applies the MSCPO technique to emergency room discharges (ERD), including both outpatients and those admitted as inpatients, for diseases of the respiratory system in South Carolina. Specifically, we monitor weekly ERD for acute upper respiratory infections (AURI), influenza, acute bronchitis, asthma, and pneumonia in 2009. The data were obtained by county for the 46 counties of South Carolina from the South Carolina Office of Research and Statistics. Total weekly ERD in South Carolina are displayed in Figure 2. The right Y axis corresponds to ERD for AURI, which are considerable larger throughout the year. In the United States, AURI are the most common acute diseases in the general population and one of the most common conditions for visiting a clinician.

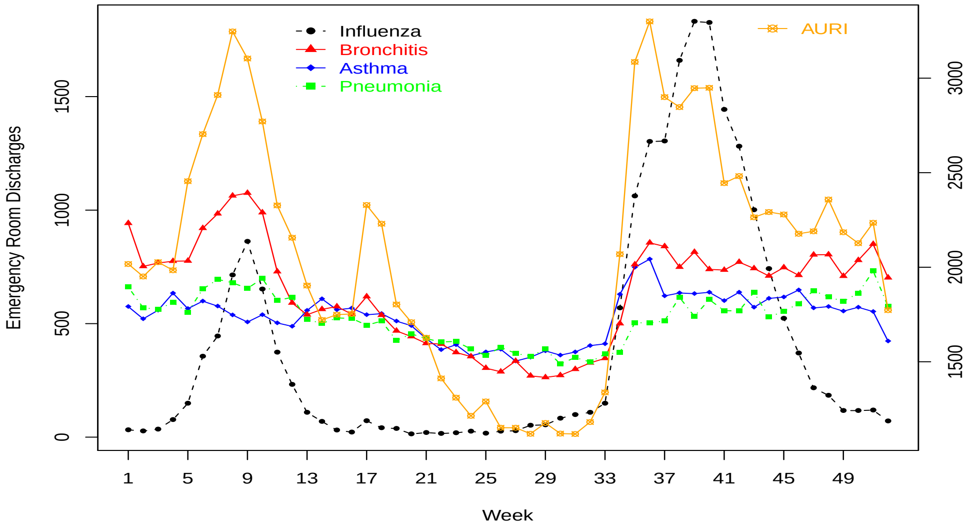

Figure 2: Weekly emergency room discharges for influenza, acute bronchitis, asthma, pneumonia, and acute upper respiratory infections (AURI, right Y axis) in South Carolina in 2009.

AURI, influenza, acute bronchitis and pneumonia are closely related acute diseases and, on some occasions, they are grouped together for data reporting, especially influenza and pneumonia. Although these diseases can happen at any time, they are most common during the fall and winter months. In the United States, peak flu season months are December, January and February. The unusual behavior shown in Figure 2 is due to the novel H1N1 influenza virus, which arrived in South Carolina in April 2009. Novel H1N1 persisted throughout the summer and the 2009-2010 influenza season, which peaked during early October and November. Asthma, on the contrary, is a chronic lung disease that inflames and narrows the airways. However, it is known that people with asthma may experience more frequent and severe asthma attacks when they have an upper respiratory infection.

Because we are interested in detecting epidemic onsets, we confine our analysis to data collected from week beginning June 28 (where all the diseases can be assumed to be in a non-epidemic state) to week beginning December 27 (weeks 26 - 52 in Figure 2). There are 46 counties, 27 time periods (weeks), and five diseases. Expected counts, which are assumed to be constant during the surveillance exercise to properly identify emerging outbreaks, were calculated for each disease and county by internal standardization[32] using the data from the first three weeks. These data were also used to initially estimate the multivariate model describing the behavior of diseases under non-epidemic conditions. Model (4) was fitted with $L = 10$ latent components. The results displayed are computed from 10000 iterations after a burn-in of 50000 iterations. Similar to the simulation study, the following prior distributions were assumed: $p_l \sim Be(1, l)$, $\delta_{l,k} \sim N(0, 100)$, and $Ga(2, 0.5)$ for the precision parameters. In this example, five spatial fields are selected. The first one is common to AURI, acute bronchitis, asthma and pneumonia, while the other four spatial fields are only relevant to one disease. Namely, they are relevant to AURI, influenza, asthma, and pneumonia, respectively. So, influenza does not share a common spatial field with the other diseases. Figure 3 displays the estimated latent spatial fields.

Table 4 shows the DIC values (together with the PD) for the estimated shared component model. For comparative purposes, we also include the DIC values for the shared component model used by Ma and Carlin[23] (Equation (3)) and those obtained when the diseases are modeled separately by using the convolution model. To select the model that best explains the correlation across both locations and diseases, the upper half of the table shows the results obtained with these models when only spatially structured random effects are incorporated into the model. The lower half of the table shows the results when both spatially correlated random effects and disease-specific spatially uncorrelated terms (residuals) are included in the model. As can be seen, the joint spatial analysis of the data with the proposed shared component model leads to an improved goodness of fit as judged by a lower overall DIC value. The model used by Ma and Carlin[23] and our model provide a similar goodness of fit when residuals are included in the model. However, by comparing the DIC values in the upper half of the table, it can be seen that a single spatial field cannot explain properly the correlation across both locations and diseases present in the data. This can be further corroborated by examining the residuals ψ_{ik} in Equation (3). The estimated residuals (not shown) present a spatial correlation, which violates the assumption about the independence of residuals.

In what follows, we show the results obtained in the prospective analysis of the data using our surveillance technique. At each time point $t = 4, 5, \ldots, 27$, the shared component model with five spatial fields is estimated using the data observed up to time $t - 1$, and the MSCPO values associated with the new observations are analyzed to detect emerging outbreaks of diseases. An alarm for the ith county is sounded at time t if the sMSCPO$_{it}$ is below $0.5 \times 10^{-n_{it}}$, n_{it} being the number of counts higher than expected in county i and time t. In order

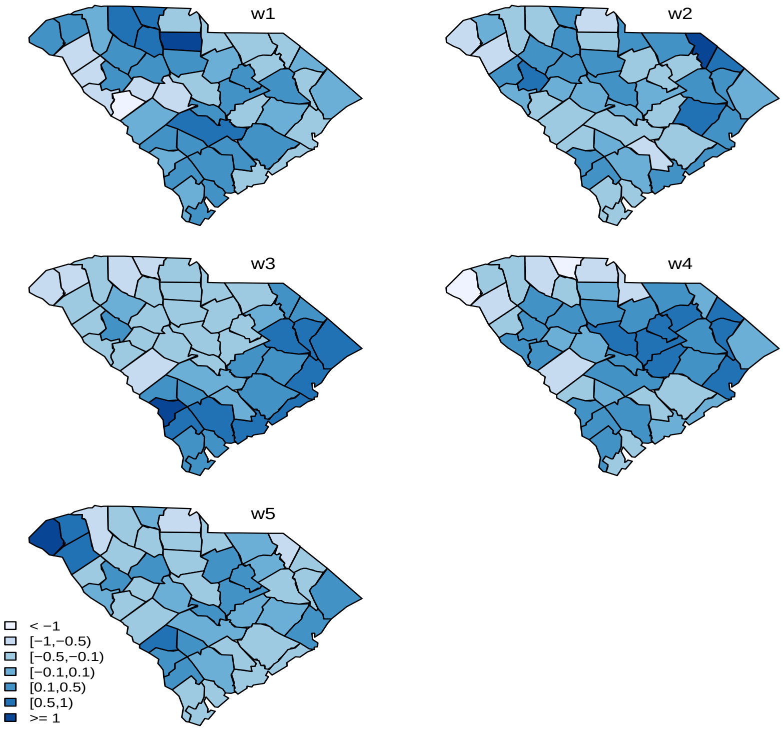

Figure 3: Case study. Estimated latent spatial fields from the shared component model. The component w1 is shared by acute upper respiratory infections (AURI), acute bronchitis, asthma and pneumonia. w2, w3, w4, and w5 are only relevant to one disease: AURI, influenza, asthma and pneumonia, respectively.

to detect not only the onset but also the end of an epidemic, counts of disease detected as unusual at time t are assumed to be missing when they become part of the history. This way, the shared component model is sequentially estimated using only data observed under non-epidemic conditions. Table 5 shows, for a selection of twenty-eight counties in South Carolina, the time point at which an outbreak is detected for each one of the diseases. Most of these outbreaks of disease are also detected when the diseases are monitored separately by using the univariate SCPO. As an example, Figures 4 and 5 show the temporal profiles for the Charleston and Greenville counties, where highlighted points represent time periods corresponding to epidemic stages. As can be seen, when observed counts of disease are unusually high in comparison with the expected counts the univariate and multivariate surveillance techniques signal an alarm at the same time. However, by borrowing information from different diseases, the MSCPO alerts us to unusual counts of disease which are not significant enough

Model	AURI	Influ	Bronch	Asthma	Pneum	Total
Shared component	810.35	268.41	583.41	659.04	651.04	2972.26
model	(38.62)	(20.42)	(27.99)	(30.98)	(26.97)	(144.97)
Ma and Carlin's	828.03	312.47	645.44	719.23	732.52	3237.69
model	(29.52)	(2.33)	(9.97)	(6.97)	(4.09)	(52.88)
Convolution model	815.37	271.61	590.30	663.03	659.08	2999.39
	(41.97)	(17.47)	(34.80)	(32.01)	(29.64)	(155.89)
Shared component	808.18	268.04	578.84	657.64	652.43	2965.13
model	(39.85)	(20.06)	(32.07)	(33.22)	(32.27)	(157.46)
Ma and Carlin's	808.90	266.51	580.71	657.80	653.84	2967.75
model	(39.94)	(18.52)	(31.20)	(31.53)	(31.42)	(152.61)
Convolution model	810.30	268.24	584.29	659.17	656.17	2978.16
	(40.80)	(19.92)	(33.88)	(33.86)	(32.62)	(161.08)

Table 4: Case study. DIC (PD) values for the shared component model, Ma and Carlin's model, and individual convolution models. Results obtained when the models only include spatially correlated random effects are shown in the upper half of the table. The lower half shows the results when disease-specific spatially uncorrelated terms (residuals) are also incorporated into the models.

to cause an alert on their own. This is the case, for instance, of the AURI, asthma and pneumonia epidemics in Greenville at the moment of their onsets or the pneumonia epidemic in Charleston, where only some extremely high observations are detected in the separate analysis of the disease.

County	AU	In	Br	As	Pn	County	AU	In	Br	As	Pn
Calhoun	9	9	10	9	9	Laurens	9	9	9	9	10
Charleston	9	5	9	9	10	Lee	10	9	10	10	10
Cherokee	7	8	8	9	12	Lexington	8	5	9	9	10
Chester	8	10	6	9	10	Marion	10	10	11	10	10
Chesterfield	10	6	9	10	10	Marlboro	8	6	9	5	5
Clarendon	9	11	-	9	13	Newberry	9	9	10	11	9
Darlington	10	10	8	10	11	Orangeburg	9	6	9	9	9
Dillon	11	8	10	10	9	Richland	9	7	7	9	9
Fairfield	9	9	10	7	9	Saluda	10	10	12	12	12
Florence	9	5	10	10	10	Spartanburg	8	8	7	8	10
Greenville	8	6	9	8	10	Sumter	9	5	9	9	9
Greenwood	5	5	9	9	8	Union	9	9	9	9	11
Kershaw	9	10	10	10	9	Williamsburg	10	10	10	10	10
Lancaster	11	6	11	6	11	York	9	9	9	9	10

Table 5: Case study. A selection of twenty-eight counties of South Carolina: Time point at which an outbreak of disease is detected based on the multivariate surveillance conditional predictive ordinate. AU: Acute upper respiratory infections; In: Influenza; Br: Acute bronchitis; As: Asthma; Pn: Pneumonia.

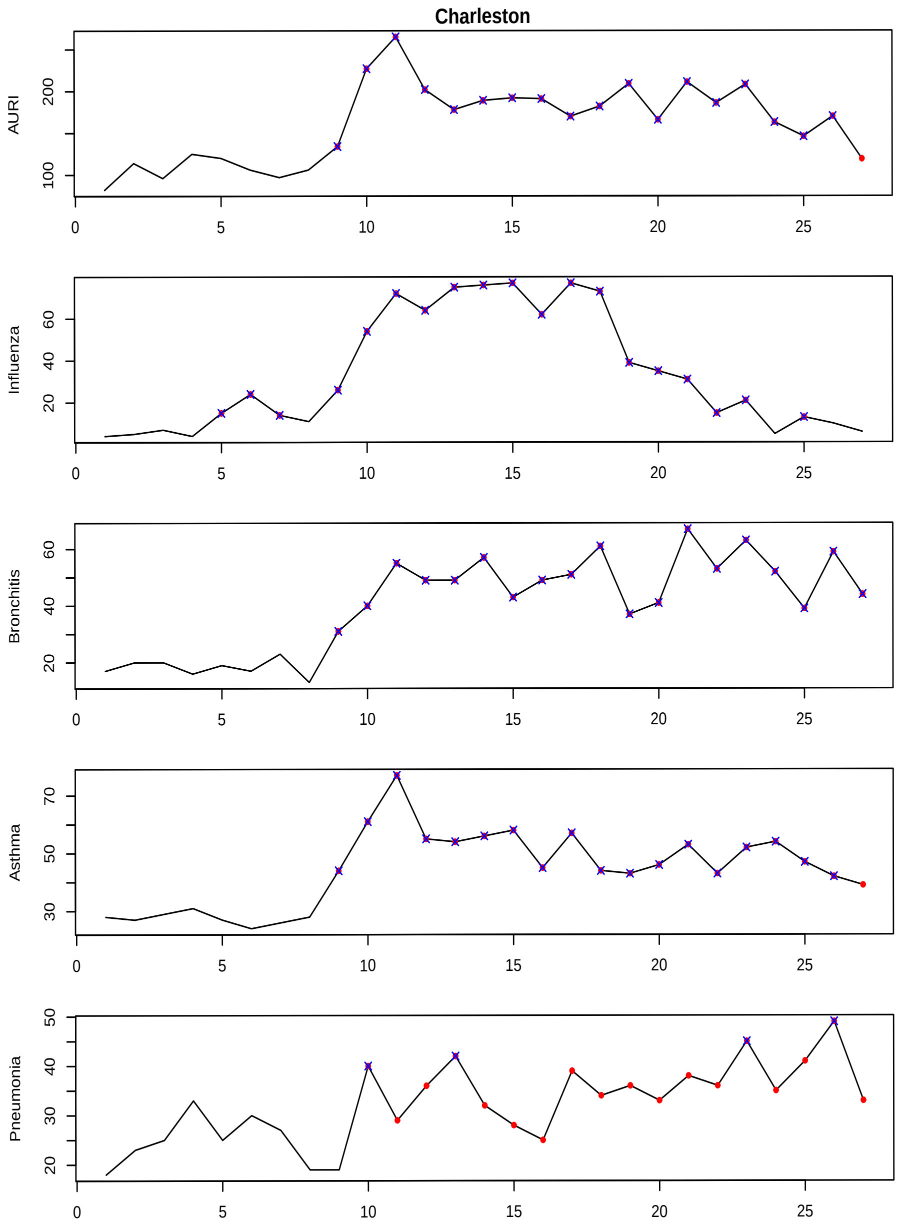

Figure 4: Temporal profile for the Charleston county. Time points corresponding to epidemic stages as detected by the multivariate surveillance conditional predictive ordinate are represented by solid points. Unusual observations based on the univariate surveillance technique are represented by crosses.

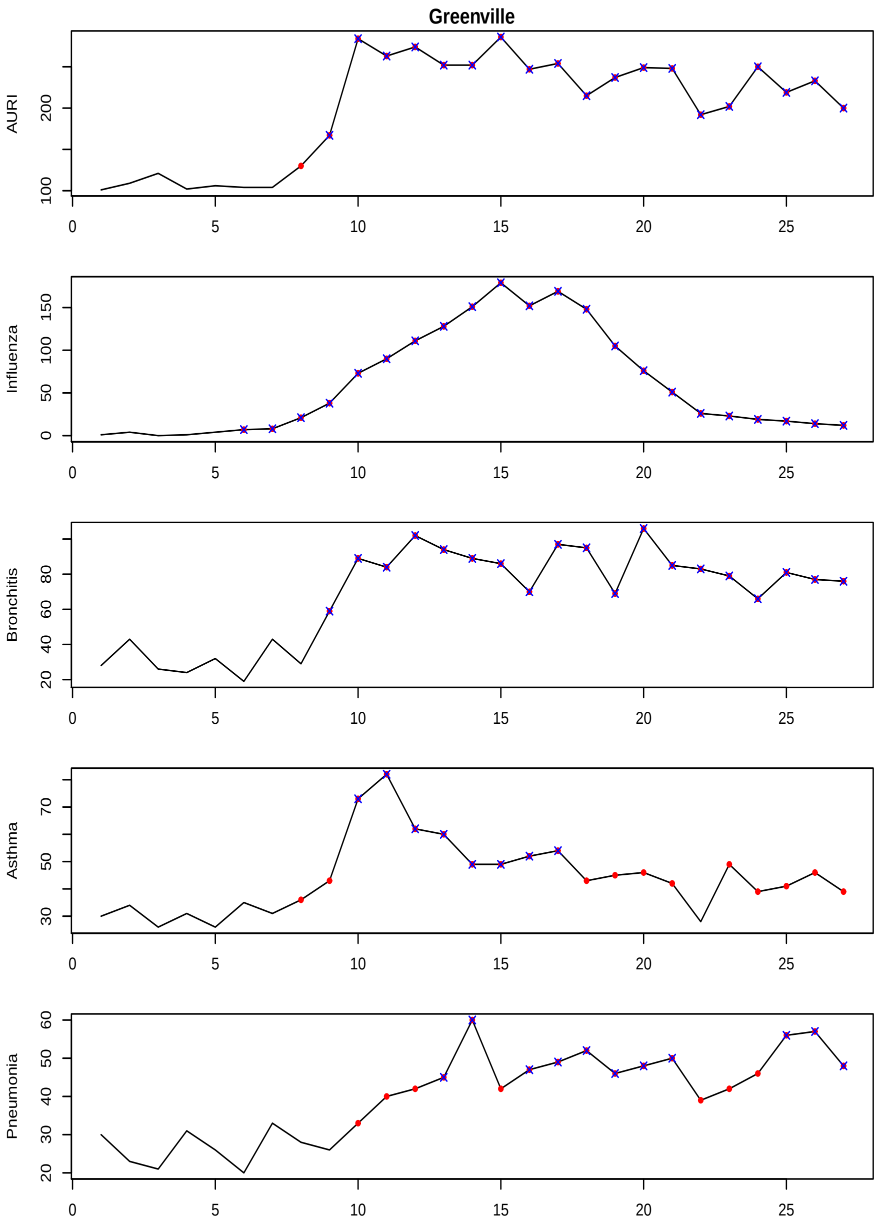

Figure 5: Temporal profile for the Greenville county. Time points corresponding to epidemic stages as detected by the multivariate surveillance conditional predictive ordinate are represented by solid points. Unusual observations based on the univariate surveillance technique are represented by crosses.

Finally, we present the results obtained with the multivariate scan statistic[11] as implemented in the free SaTScan$^{\text{TM}}$ software[33]. This method is an extension of the space-time scan statistic[4] with the ability to detect clusters in either one or in a combination of data sets. Here, the Poisson-based prospective space-time scan statistic is used. We set the maximum spatial cluster size at 50% of the population at risk, which is the default setting, and the maximum temporal window size at 90% of the study period. The non-parametric spatial adjustment provided by the software is used to adjust for purely spatial clusters. In this example, the first alarm is sounded at time period 7. In addition to the most likely cluster, which includes seven counties in the northwest of South Carolina, a statistically significant secondary cluster is detected. The criterion for reporting secondary clusters used here is that no cluster centers are included in other clusters. Table 6 shows a summary of the results provided by the software at this time period.

1. *Most likely cluster*: Spartanburg, Greenville, Cherokee, Union, Laurens, Pickens, Anderson

Time frame: 2009/08/09 to 2009/08/15 (week 7 in the analysis)
p-value: 0.0001

	Cases	Expected	Relative risk
AURI	354	327.25	1.09
Influenza	16	9.50	1.72
Bronchitis	136	96.00	1.47
Asthma	100	85.50	1.18
Pneumonia	121	98.25	1.25

2. *Secondary cluster*: Clarendon, Sumter, Williamsburg, Calhoun, Lee, Orangeburg, Florence, Berkeley, Dorchester, Darlington, Kershaw, Richland, Georgetown, Bamberg, Marion, Colleton, Lexington, Charleston, Chesterfield, Dillon, Fairfield, Barnwell, Marlboro

Time frame: 2009/08/02 to 2009/08/15 (weeks 6-7 in the analysis)
p-value: 0.04

	Cases	Expected	Relative risk
Influenza	143	108.50	1.54
Bronchitis	236	220.50	1.09

Table 6: Case study. Clusters detected by the multivariate space-time scan statistic at time period 7.

Most of the counties included in these two clusters are also detected with the MSCPO which, as shown in Table 5, signals the first alarm at time 5. However, there are some differences between these two procedures that are worthy to emphasize. The space-time scan statistic pinpoints the general time and location of the most likely cluster (and possible significant secondary clusters), and so its exact boundaries remain uncertain. This cluster corresponds to the cylinder with the maximum likelihood ratio. As a consequence, areas with no increase in the number of cases reported can be included in the cluster if its neighbors present an increased disease incidence. This is the case, for instance, of the Union county, where the observed counts of disease at time 7 are similar to those observed at previous time periods. It is also possible that the counties included in the cluster do not undergo an outbreak of disease for all the diseases reported in the cluster. For instance, only the number of ERD for influenza presents an increase in the Greenville county at time 7 (see Figure 5). Finally, because the scan statistic focuses on the detection of the most likely cluster (secondary clusters), small outbreaks of disease may be missed or reported at later time periods. These conclusions apply to all the time periods during the surveillance exercise. As an example, Figure 6 compares the counties declared as epidemic areas based on the MSCPO and the multivariate scan statistic at six time periods. As can be seen, by detecting at each time point those areas of increased disease incidence and the diseases within each area with more counts than expected, our surveillance technique enables more accurate outbreak detection and, consequently, a timelier and more informed response.

6 Discussion

The SCPO was introduced in a univariate surveillance setting to monitor spatially aggregated disease incidence data. The surveillance technique generates an alarm for the ith small area at time period t if the conditional predictive distribution of the new count of disease given the data collected so far is below a critical level α, which controls the trade-off between false alarms and detection delay or detection probability. To assure a low probability of false alarm, a small value of α should be considered. Consequently, small increases in disease incidence may be missed. The results from a simulation study and the subsequent application to emergency room discharges for five diseases of the respiratory system demonstrate that, by integrating information from multiple diseases, the multivariate surveillance technique proposed in this paper achieves substantial improvements in both detection time and recovery of the true outbreak behavior when changes in disease incidence happen simultaneously for two or more diseases.

Since the MSCPO does not depend on the model describing the behavior of diseases under non-epidemic conditions, it can be applied in any surveillance context where a statistical model is used to describe spatial data on multiple diseases. We have focused here on Bayesian hierarchical Poisson models. In particular, we have proposed a new shared component model formulation that uses binary indicator variables to identify shared and disease-specific spatially correlated latent fields. Joint modeling improves relative risk estimation and goodness of fit when the diseases under study are influenced by common confounding factor. In practice, however, this is not always the case, and so the

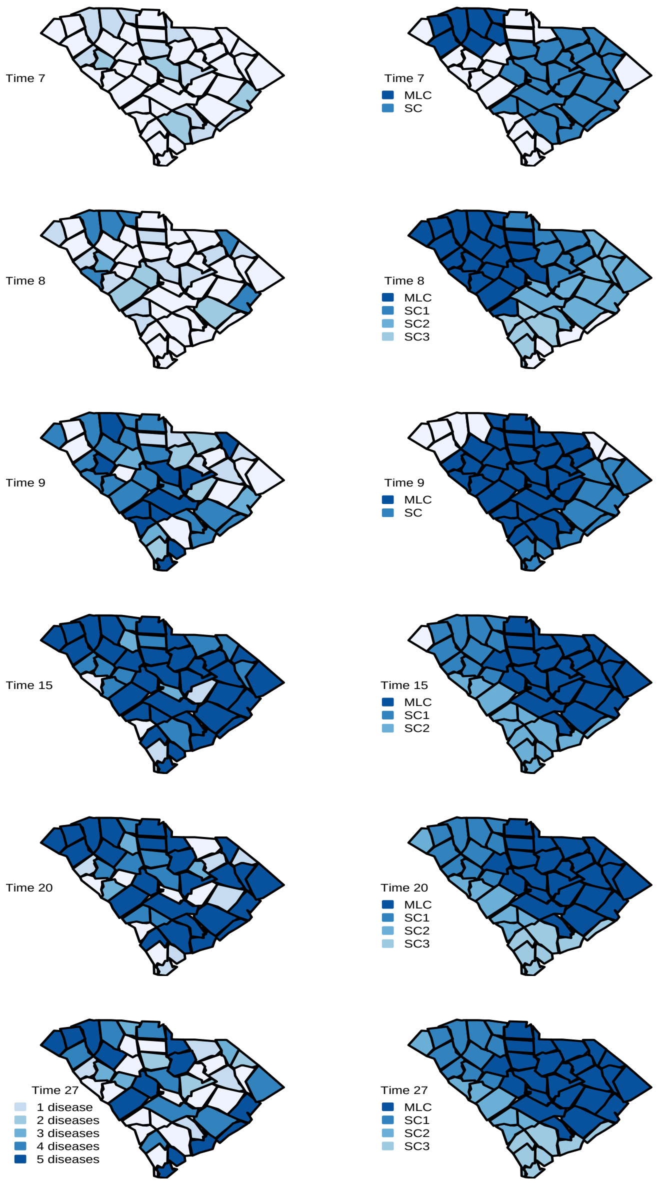

Figure 6: Case study. Counties declared as epidemic areas. Left: Areas signaling an alarm based on the multivariate surveillance conditional predictive ordinate. Darker shading indicates a higher number of diseases causing the alarm. Right: Most likely cluster (MLC) and secondary clusters (SC) using the Poisson-based prospective space-time scan statistic.

model includes only disease-specific spatial fields when the diseases of interest are independent. This is equivalent to fitting separate convolution models. A well-known problem with latent structure models is identifiability of the latent components. Empirical evidence of identification is apparent in both the simulated data and the case study. However, additional restrictions such as orthogonality of the latent components may be necessary on some occasions.

As mentioned before, our interest in this paper has been to propose a multivariate surveillance technique to jointly monitor multiple diseases in an effort to detect epidemics at the very moment of their onset. Here the diseases are assumed to be equally important. However, it may be the case that some of the diseases under study have a special relevance. For instance, in the case study, epidemics of more serious diseases of the respiratory system, such as bronchitis and pneumonia, may be particularly important. As we have shown, these epidemics are usually preceded by epidemics of milder diseases such as AURI or influenza. It would be valuable to investigate how this information can be used to predict changes in the relative risk pattern of the diseases of interest. This line of research is particularly useful in a syndromic surveillance setting, where information regarding syndrome-based outbreaks can be used to predict increases in the incidence of the disease of interest.

References

[1] Thacker SB. Historical Development. In: Lee LM, Teutsch SM, Thacker SB and Louis ME (eds) *Principles & Practice of Public Health Surveillance.* 3rd ed. Oxford: Oxford University Press, 2010, pp.1-17.

[2] Unkel S, Farrington CP, Garthwaite PH, Robertson C and Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society, series A* 2012; 175: 1-34.

[3] Rogerson PA and Yamada I. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine* 2004; 23: 2195-2214.

[4] Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, series A* 2001; 164: 61-72.

[5] Kleinman K, Lazarus R and Platt R. A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism. *American Journal of Epidemiology* 2004; 159: 217-224.

[6] Diggle P, Rowlingson B and Su T. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 2005; 16: 423-434.

[7] Vidal Rodeiro CL and Lawson AB. Monitoring Changes in Spatio-temporal Maps of Disease. *Biometrical Journal* 2006; 48: 463-480.

[8] Zhou H and Lawson AB. EWMA smoothing and Bayesian spatial modeling for health surveillance. *Statistics in Medicine* 2008; 27: 5907-5928.

[9] Watkins RE, Eagleson S, Veenendaal B, Wright G and Plant AJ. Disease surveillance using a hidden Markov model. *BMC Medical Informatics and Decision Making* 2009; 9: 39.

[10] Robertson C, Nelson TA, MacNab YC and Lawson AB. Review of methods for space-time disease surveillance. *Spatial and Spatio-temporal Epidemiology* 2010; 1: 105-116.

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

[12] Neill DB and Cooper GF. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 2010; 79: 261-282.

[13] Banks D, Datta G, Karr A, Lynch J, Niemi J and Vera F. Bayesian CAR models for syndromic surveillance on multiple data streams: Theory and practice. *Information Fusion* 2010; doi:10.1016/j.inffus.2009.10.005.

[14] Gelfand AE and Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 2003; 4: 11-25.

[15] Knorr-Held L and Best NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, series A* 2001; 164: 73-85.

[16] Held L, Natário I, Fenton SE, Rue H and Becker N. Towards joint disease mapping. *Statistical Methods in Medical Research* 2005; 14: 61-82.

[17] Corberán-Vallet A and Lawson AB. Conditional predictive inference for online surveillance of spatial disease incidence. *Statistics in Medicine* 2011; 30: 3095-3116.

[18] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* 2000; 19: 2555-2567.

[19] Lawson AB. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology.* Boca Raton: Chapman & Hall, 2009.

[20] Lawson AB. Spatial and Spatio-Temporal Disease Analysis. In: Lawson AB and Kleinman K (eds) *Spatial and Syndromic Surveillance for Public Health.* Chichester: Wiley, 2005, pp.55-76.

[21] Besag J, York J and Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; 43: 1-59.

[22] MacNab YC. On Bayesian shared component disease mapping and ecological regression with errors in covariates. *Statistics in Medicine* 2010; 29: 1239-1249.

[23] Ma H and Carlin BP. Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis. *Bayesian Analysis* 2007; 2: 281-302.

[24] Wang F and Wall MM. Generalized common spatial factor model. *Biostatistics* 2003; 4: 569-582.

[25] Oleson JJ, Smith BJ and Kim H. Joint Spatio-Temporal Modeling of Low Incidence Cancers Sharing Common Risk Factors. *Journal of Data Science* 2008; 6: 105-123.

[26] Lawson AB, Song HR, Cai B, Hossain MM and Huang K. Space-time latent component modeling of geo-referenced health data. *Statistics in Medicine* 2010; 29: 2012-2027.

[27] Frühwirth-Schnatter S and Lopes HF. Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical report, University of Chicago Booth School of Business, 2009.

[28] Congdon P. *Bayesian Statistical Modelling.* 2nd ed. Chichester: John Wiley & Sons, 2006, pp.425-455.

[29] Geisser S. Discussion on Sampling and Bayes' Inference in Scientific Modelling and Robustness (by GEP Box). *Journal of the Royal Statistical Society, series A* 1980; 143: 416-417.

[30] Gelfand AE, Dey DK and Chang H. Model determination using predictive distributions with implementation via sampling-based methods. In: Bernardo JM, Berger JO, Dawid AP and Smith AFM (eds) *Bayesian Statistics 4*. Oxford: Oxford University Press, 1992, pp.147-167.

[31] Congdon P. *Bayesian Models for Categorical Data.* Chichester: John Wiley & Sons, 2005, pp.39-40.

[32] Banerjee S, Carlin BP and Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton: Chapman & Hall, 2004, pp.153.

[33] SaTScan$^{\text{TM}}$ v9.1.1: Software for the spatial and space-time scan statistcs, 2009 (http://www.satscan.org/). Accessed 26 October 2011.