Published in final edited form as:

Biometrics. 2011 December ; 67(4): 1508–1517. doi:10.1111/j.1541-0420.2011.01585.x.

Detecting Disease Outbreaks Using Local Spatiotemporal Methods

Yingqi Zhao¹, Donglin Zeng¹, Amy H. Herring², Amy Ising³, Anna Waller³, David Richardson⁴, and Michael R. Kosorok^{1,*}

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

²Department of Biostatistics, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

³Carolina Center for Health Informatics, Department of Emergency Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

⁴Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

Summary

A real-time surveillance method is developed with emphasis on rapid and accurate detection of emerging outbreaks. We develop a model with relatively weak assumptions regarding the latent processes generating the observed data, ensuring a robust prediction of the spatiotemporal incidence surface. Estimation occurs via a local linear fitting combined with day-of-week effects, where spatial smoothing is handled by a novel distance metric that adjusts for population density. Detection of emerging outbreaks is carried out via residual analysis. Both daily residuals and AR model-based de-trended residuals are used for detecting abnormalities in the data given that either a large daily residual or an increasing temporal trend in the residuals signals a potential outbreak, with the threshold for statistical significance determined using a resampling approach.

Keywords

Disease surveillance; Local linear estimation; Residual analysis; Lattice Data; Time series modeling

1. Introduction

The primary purpose of disease surveillance is to detect unusual spatial or temporal patterns of disease; this may lead to further investigation to determine the causes of abnormal events. To provide early warning and enable rapid public health intervention, it is important to develop a reliable disease surveillance system that can quickly identify anomalies. Such a system might combine available information sources such as emergency department (ED) visit data, physician office data, disease reports from clinical laboratories, and data on over-the-counter drug sales.

The North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) was created by the NC Division of Public Health (NC DPH) in 2004 in

*kosorok@unc.edu.

collaboration with the UNC Department of Emergency Medicine to provide statewide early recognition of outbreaks and monitoring of public health using various data sources. Currently, NC DETECT's sources of data include emergency departments (ED), the Carolinas Poison Center, and the Pre-hospital Medical Information System (PreMIS), as well as pilot data from the NCSU College of Veterinary Medicine Laboratories and select urgent care centers. All hospitals in NC with 24-hour-acute care emergency departments must report electronic data in near-real time. This paper focuses on the use of spatiotemporal methods for early identification and situational awareness of unexpected variation in the incidence of asthma reported to NC DETECT from emergency departments, considering ED admissions that occurred between January 1, 2006 and December 31, 2008.

A variety of methods have been developed for the purpose of surveillance. A standard surveillance tool used by CDC, as well as NC DETECT, is cumulative sums (CUSUM), a quality control method (Hutwagner et al. (1997)). CUSUM accumulates deviations between observed and expected values; if the expectation is modeled poorly, signals detected by CUSUM may not reflect the true underlying changes. Time series approaches are also used to detect disease outbreaks (Reis and Mandl (2003), Craigmile et al. (2007)), although these methods do not account for spatial correlation. Openshaw et al. (1987) developed a graphical method, the geographical analysis machine (GAM), counting the observed cases in multiple overlapping circles with increasing radii and identifying those significantly higher than the expected value. GAM, however, is criticized by Kulldorff and Nagarwalla (1995) for its limitation in handling multiple testing. Kulldorff (1997) proposed a spatial scan statistic, widely used for cluster investigations, which searches over a given set of regions and applies the likelihood ratio test for the null hypothesis that the probability of events happening in the window is the same as that outside the window, while the alternative hypothesis represents clustering in that region, the underlying disease rate is higher inside than outside. However, it is difficult to derive the exact distribution of this likelihood ratio statistic. Bernoulli and Poisson data are analyzed using the spatial scan statistic in Kulldorff (1997), and it has also been generalized to ordinal data (Jung et al. (2007)), and continuous data (Kulldorff et al. (2009)). Cucala, Demattei, Lopes, and Ribeiro (Cucala et al.) note that the spatial scan statistic can be computationally infeasible when the number of locations is too large, proposing instead graph-based spatial scan tests linking those events closer than a given distance, allowing completely data-based clusters rather than only those of predetermined shape. This has been generalized to the detection of disease clusters in the space-time domain (Kulldorff et al. (1998)). Application of prospective disease surveillance may be done using space-time scan statistics to detect new clusters resulting from an emerging disease outbreak (Kulldorff (2001), Neill et al. (2005)).

Representing the cases as a point pattern, it is natural to extend point process methodology to the surveillance problem. There is a vast literature on spatial point processes, motivated by growing numbers of data sets being collected in fields such as epidemiology, environmental studies, geography, seismology and forestry (Ripley (1977), Cressie (1993), Møller and Waagepetersen (2004)). However, methodology for the analysis of spatiotemporal point processes is less well developed.

One common approach for the analysis of spatiotemporal point processes is to consider the conditional intensity, which can be interpreted as the hazard of the occurrence of an event at location s and time t , given the history of the process over the interval $[0, t]$, which uniquely determines the probability structure of the point process when it exists (see Daley and Vere-Jones (2003)).

We will obtain lattice data, also known as areal data, if cases are aggregated over geographical regions such as counties. General inferential issues for lattice data focus on

models for spatial clustering and variation. Traditionally, a neighborhood structure can be defined to introduce spatial correlation, and such data are modeled by borrowing information from neighboring regions. Two pervasive models that incorporate such neighborhood information are the simultaneously and conditionally autoregressive models (SAR and CAR), pioneered by Whittle (1954) and Besag (1974), respectively.

Diggle et al. (2005) formulate the problem of online spatiotemporal disease surveillance by obtaining predictions in the context of a non-stationary log-Gaussian Cox process and calculating the exceedance probabilities of the intensities over a pre-specified threshold value. Unfortunately, unreliable estimates may result from violations of the Cox model assumption.

We develop a real-time surveillance method, utilizing a local linear estimation method incorporating day-of week effects to construct the predicted spatiotemporal incidence rate surfaces and implementing residual analysis to identify anomalies. In using a model-based approach, we attempt to model the available data from a normal period, and then identify a future emerging aberration as quickly as possible. We do not impose Poisson process or log Gaussian Cox process (Diggle et al. (2005)) assumptions where intensity functions are straightforward for modeling; instead, we only assume event counts at each location follow a Poisson distribution with a marginal Poisson rate. This makes fewer assumptions about the latent processes, in the hope of being able to ensure greater robustness in prediction than existing methods. We first form a model to estimate marginal Poisson rates, which is complex enough to capture the spatiotemporal structure of the data; the proposed prediction methodology accounting for spatial and temporal variation in the underlying incidence rates enables us to detect true clusters more precisely. We can then carry out flexible testing for abnormalities based on residual analysis in a computationally fast manner. With this approach, we construct a model describing normal circumstances, which can inform public health administrators of expected events in addition to alarms. Thus the surveillance system we develop has usefulness beyond cluster detection because users can gain information on the regular pattern of incidence, which other methods (e.g. scan statistics) cannot provide.

In Section 2 we propose a spatiotemporal model for the marginal Poisson rate estimated using local linear regression methods. We then discuss detection of outbreaks via residual analysis. In Section 3 we present a data application using asthma ED admissions to illustrate how the proposed surveillance system works and provide results of a carefully designed simulation study comparing our method to the CUSUM and space-time scan statistic (Kulldorff (2001)). Finally, we outline future research and open questions in Section 4.

2. Methodology

2.1 Model

Let the number of observed cases at location s on day t be $N(s, t)$, and let $\pi(s)$ be the population of location s , which we assume is constant during the study period. We want to provide a reliable prediction of the normal pattern of spatial and temporal incidence of cases, i.e., we need to estimate the incidence rates for the underlying spatiotemporal point process from the collected data.

When studying events such as hospital admissions or emergency department visits, there are some commonly-observed systematic patterns to event occurrence. For example, the number of ED admissions at a hospital tends to be greater on weekends than on week days. Therefore, natural day-of-week variation must be accounted for in the model. We assume that the observed incidence rate at a given location and on a given day is the product of a

base incidence rate at this location and a value dependent on the day of week. The resulting model for the Poisson rate $\lambda(s, t)$ is:

$$\lambda(s, t) = \mu(s, w) e^{\sum_{l=1}^7 \alpha_l(s) I(\text{Day}=l)} + \epsilon(s, t), \quad (1)$$

subject to

$$\sum_{l=1}^7 \alpha_l(s) = 0$$

where $\mu(s, w)$ is the baseline Poisson rate of location s in the week w in which day t , the l^{th} day of the week, occurs (i.e. for Sunday, January 8, $t = 8$, $w = 2$, $l = 1$). We assume the day-of-week effect, denoted by $\alpha_l(s)$, $l = 1, \dots, 7$, is the same across the whole period but varies across the region, allowing each location to have a specific weekly pattern. Then we estimate the expected rate at location s on day t and use this to detect the outbreaks.

2.2 Estimation

The model (1) has two types of parameters: local parameters for the weekly baseline effect $\mu(s, w)$ and global parameters for the day-of-week effect $\alpha_l(s)$. Local parameters are only determined by a local neighborhood, while global parameters depend on all locations. We discuss the estimation procedures in the following sections.

2.2.1 Local Linear Estimates for Baseline Weekly Effect—The weekly baseline rate is the geometric mean of each day's rate in the given week. We focus on the smoothing method for incidence rates by day first and then discuss the estimation of day-of-week effects.

Nonparametric smoothing methods have been widely used in many statistical areas, although they have seen relatively little use in the point process field. Diggle et al. (1985) discusses kernel smoothing methods for one-dimensional point processes. The weaknesses of this type of kernel smoother have been well discussed, including its poor boundary performance, large bias and low efficiency (see e.g. Fan and Gijbels (1996)). The boundary effects can be even worse in two or higher dimensions. We therefore smooth the three-dimensional spatiotemporal point processes via local linear fitting, which is gaining increasing popularity for its desirable properties over kernel smoothing.

We implement local linear fitting by approximating the unknown Poisson rate function λ by a linear function in time in the neighborhood of any point (s, t) , using weighted least squares, with weights provided by a spatiotemporal kernel. For location s , we estimate $\lambda(s, w_k)_l$, $l = 1, \dots, 7$, which is the rate in location s for the l^{th} day of the k^{th} week, denoted by w_k , in a year, based on that day's data (e.g. all Mondays).

For every l^{th} day of the week, at each (s, t_l) , we solve

$$\min_{\beta_0, \beta_1} \sum_{i,k} K_{i,k}(s, t_l) (\lambda(s_i, t_{w_{kl}}) - \beta_0(s, t_l) - \beta_1(s, t_l)(t_{w_{kl}} - t_l))^2, \quad (2)$$

for $\hat{\beta}_0(s, t_l)$, $\hat{\beta}_1(s, t_l)$. Here i indexes the location, $i = 1, \dots, S$, and k indexes the weeks, $k = 1, \dots, M$. t_l represents the l^{th} day of a week generally, $\lambda(s_i, t_{w_{kl}}) = N(s_i, t_{w_{kl}})/n(s_i)$, $N(s_i, t_{w_{kl}})$ is the recorded visit count on the l^{th} day of k^{th} week, denoted by $t_{w_{kl}}$ at location s_i and $n(s_i)$ is the population of location s_i . $K(s, t_l)$ is the kernel function assigning weights to

points in a region around (s, t) based on the distance from (s, t) with effect from (s_i, t_{wkl}) calculated as

$$K_{i,k}(s, t) = K_1 \left(\frac{\|s - s_i\|}{c_{s,t} a_n} \right) K_2 \left(\frac{\|\theta(t) - \theta(t_{wkl})\|}{a_n} \right). \quad (3)$$

This kernel function is factored into a spatial kernel K_1 and a temporal kernel K_2 , accounting for spatial and temporal effects on the incidence rate. We use a Gaussian kernel function for both kernels. The weight decreases exponentially with the distance in the spatial domain from s in K_1 and in the temporal domain from t in K_2 . The temporal distance is defined based on the function $\theta(t) = \exp(i2\pi t/365)$. We also define a new distance metric—population density adjusted distance—in the space kernel, and we leave this for later discussion.

a_n in the kernel function is a sequence of bandwidths tending to zero as n goes to infinity. Note that a_n controls the size of the local temporal neighborhood. Because of different scales in space and time, we use a scaling factor $c_{s,t}$ in the kernel K_1 , where

$$c_{s,t} = \frac{\text{median}_i \|s - s_i\|}{\text{median}_k \|\theta(t) - \theta(t_{wkl})\|}.$$

This simplifies the selection of two bandwidths simultaneously in a reasonable way because $c_{s,t}$ reflects the relative magnitude in the spatial domain compared to the temporal domain. Demattei and Cucala (2011) introduce a weighted combination of the spatial and temporal Euclidian distances with a factor handling correspondence between space and time. Their spatio-temporal distance measure is based on the addition of two distances after rescaling the temporal axis, while we allow for the effect on spatial and temporal neighborhoods to vary separately, considering that the information is more balanced in this context. By multiplying a_n by $c_{s,t}$ we can have the bandwidth controlling the size of the spatial neighborhood. The bandwidths can then be selected via data-driven cross validation.

The solution to the minimization problem (2) is easily shown to be $(X'WX)^{-1}X'W\Lambda$, where Λ is a vector of length ST with stacked rates, and W is an $ST \times ST$ diagonal matrix with $((7k + l - 1)S + i)^{th}$ entry being $K_{i,k}(s, t)$, where T is the total number days during the study period. Thus we obtain

$$\widehat{\lambda}(s, t) = \widehat{\beta}_0(s, t). \quad (4)$$

Using all the estimated daily incidence rates, we can now calculate baseline weekly effects as

$$\widehat{\mu}(s, w_k) = \sqrt[7]{\prod_{l=1}^7 \widehat{\lambda}(s, t_{wkl})}, \quad (5)$$

where $\widehat{\lambda}(s, t_{wkl})$ is obtained from (4).

2.2.2 Temporal Distance—For two time points t_i, t_{wkl} the length of the time interval $|t_i - t_{wkl}|$ is a common distance measure. We propose instead to use the temporal distance $|\theta(t_i) - \theta(t_{wkl})|$, where $\theta(t) = \exp(i2\pi t/365)$, ensuring the rates at the end of one year and beginning of the next should be close, as this complex number distance metric has attractive features in representing cyclical patterns.

2.2.3 Population density adjusted distance—Distance measures widely used in spatial data analysis include ordinary Euclidean distance, great circle distance, and others. We are interested in the disease cluster being adjusted for spatial variations in the population density because potential clustering can emerge by chance due solely to population clustering. There is information about case incidence in the population, and intuitively, more densely populated areas contain more information than rural areas. For the spatial distance between two locations s and s_i ($\|s - s_i\|$ in (3)), we propose a new measure, $\rho_p(s, s_i)$, which will be defined later, for spatial smoothing that adjusts for population density in the sense that we try to stretch out the denser area and make sparser areas more compact, in order to ensure a similar degree of smoothing for all individuals in the population.

Let D_i denote the population density in the i^{th} location. Let ρ_g be the Euclidean distance measure. We propose an adjusted distance between two locations s_1, s_2 defined as

$$\rho_p(s_1, s_2) = \sum_{i=1}^S \int_0^1 \tilde{K}\left(\frac{G_i(u)}{h}\right) D_i |g_i(u)| du, \quad (6)$$

where h denotes the bandwidth of $\tilde{K}$, $G_i(u) = \rho_g(s_1 + u(s_2 - s_1), s_i)$ and $g_i(u) = dG_i(u)/du$. As u ranges from 0 to 1, $G_i(u)$ calculates the distance between s_i and any locations that lie between s_1 and s_2 . The intuition behind this adjusted distance, defined on the population density domain instead of on geographical space, is that two locations are likely to experience similar rates of disease if their population densities and structures are close. It is thus obtained by population density aggregation along the segment linking s_1 and s_2 . For any given location s , all the counties will contribute to its population density, with the influences based on their geographic distances from s . Specifically, for county s_i with density D_i , its contribution to the population density at location s is $\tilde{K}(\rho_g(s, s_i)/h)D_i$. Thus on any infinitesimal unit of the segment linking s_1 and s_2 , say $[s, s + ds]$, the population density adjusted distance for this interval contributed by county s_i is equal to $\tilde{K}(\rho_g(s, s_i)/h)D_i/d\rho_g(s, s_i)$. As a result, the adjusted distance between s_1 and s_2 is

$$\int_{s \in [s_1, s_2]} \sum_{i=1}^n \tilde{K}(\rho_g(s, s_i)/h) D_i |d\rho_g(s, s_i)|,$$

which is equivalent to (6) after transforming $[s_1, s_2]$ to $[0, 1]$.

Here we use a kernel smoothing method with a Gaussian kernel $\tilde{K}$. Assume that $G_i(u)$ is minimized globally at u_i . Define weights η_i as follows: if $u_i \in [0, 1]$,

$$\eta_i = h \left(\Phi\left(\frac{G_i(0)}{h}\right) + \Phi\left(\frac{G_i(1)}{h}\right) - 2\Phi\left(\frac{G_i(u_i)}{h}\right) \right);$$

otherwise if $u_i \notin [0, 1]$,

$$\eta_i = h \left(\Phi\left(\frac{\max\{G_i(0), G_i(1)\}}{h}\right) - \Phi\left(\frac{\min\{G_i(0), G_i(1)\}}{h}\right) \right).$$

Here Φ is the standard Gaussian cumulative distribution function. Thus we have

$$\rho_p(s_1, s_2) = \sum_{i=1}^S \eta_i D_i.$$

The proposed distance between two locations $\rho_p(s_1, s_2)$ is actually a weighted sum of all the locations' population densities, while the weights depend on the real geo-distances between s_1 and s_2 . Compared to the geo-distances commonly used in spatial statistics, the adjusted distance has appealing properties in terms of spatial smoothing in surveillance applications. We will return to this definition with an application to the ED visit data from North Carolina in Section 3.1.

2.2.4 Estimation of Day-of-Week Effect—The day-of-week effect is handled in an ANOVA-type framework. Given that we have the estimated baseline effects $\widehat{\mu}(s, w_k)$ from (5), we can use these in our objective function (1). Let $\alpha(s_i) = (\alpha_1(s_i), \dots, \alpha_7(s_i))'$, and $U_{w_k}^l = (I(t_{w_{kl}} = \text{Sun}), \dots, I(t_{w_{kl}} = \text{Sat}))'$. Then

$$\widehat{\alpha}(s_i) = \arg \min_{i,k,l} \sum \left(\lambda(s_i, t_{w_{kl}}) - \widehat{\mu}(s_i, w_k) e^{\alpha(s_i)^T U_{w_k}^l} \right)^2.$$

We can estimate $\widehat{\alpha}(s_i)$ as follows. Letting

$$Q(\alpha(s_i)) = \sum_{i,k,l} \left(\lambda(s_i, t_{w_{kl}}) - \widehat{\mu}(s_i, w_k) e^{\alpha(s_i)^T U_{w_k}^l} \right)^2,$$

we have

$$\begin{aligned} \frac{\partial Q(\alpha(s_i))}{\partial \alpha(s_i)} &= -2 \sum_{i,j} U_{w_k}^l \widehat{\mu}(s_i, w_k) e^{\alpha(s_i)^T U_{w_k}^l} \left(\lambda(s_i, t_{w_{kl}}) - \widehat{\mu}(s_i, w_k) e^{\alpha(s_i)^T U_{w_k}^l} \right), \\ \frac{\partial^2 Q(\alpha(s_i))}{\partial \alpha(s_i)^2} &= -2 \sum_{i,k,l} U_{w_k}^l U_{w_k}^{lT} \widehat{\mu}(s_i, w_k) e^{\alpha(s_i)^T U_{w_k}^l} \lambda(s_i, t_{w_{kl}}) + 4 \sum_{i,k,l} U_{w_k}^l U_{w_k}^{lT} \widehat{\mu}(s_i, w_k)^2 e^{2\alpha(s_i)^T U_{w_k}^l}, \end{aligned}$$

subject to

$$\sum_{l=1}^7 \alpha_l(s) = 0.$$

By using the Fisher scoring algorithm, we replace the observed information by its expectation, where

$$E \left(\frac{\partial^2 Q(\alpha(s_i))}{\partial \alpha(s_i)^2} \right) = 2 \sum_{i,k,l} U_{w_k}^l U_{w_k}^{lT} \widehat{\mu}(s_i, w_k)^2 e^{2\alpha(s_i)^T U_{w_k}^l}.$$

While the empirical information matrix may not always be nonnegative definite, the expected information applied in the Fisher scoring algorithm is always positive definite and can lead to a better estimate of the asymptotic covariance. Then we obtain

$$\alpha(s_i)^{n+1} = \alpha(s_i)^n + \left\{ E \left(-\frac{\partial^2 Q(\alpha(s_i)^n)}{\partial \alpha(s_i)^2} \right) \right\}^{-1} \left\{ \frac{\partial Q(\alpha(s_i)^n)}{\partial \alpha(s_i)} \right\}.$$

2.3 Detecting Outbreaks via Residual Analysis

Once we observe new data, i.e., the case records from each location, we calculate the residuals for detection of abnormalities in the data. A square root transformation is applied to the count data to stabilize the variance, with the residual for the disease rate defined as

$$X_{ij} = 2 \left(\sqrt{\frac{N(s_i, t_j)}{n(s_i)}} - \sqrt{\hat{\lambda}(s_i, t_j)} \right),$$

where j indexes the days in the study period, $j = 1, \dots, T$. This raw residual is taken after removing the systematic pattern in the counts, adjusted for spatial effects.

Each location has a residual time series of high autocorrelation. Fitting an AR(2) model to the i^{th} residual series, we have

$$X_{ij} = \mu_i + \phi_{i1}X_{i,j-1} + \phi_{i2}X_{i,j-2} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_i^2). \quad (7)$$

Our diagnostic shows that the autocorrelation vanishes after AR(2) model fitting in some degree for our data, and the residuals from the model are approximately white noise,

$\epsilon_{ij} \sim N(0, \sigma_i^2)$. Moreover, the model residual ϵ_{ij} is taken after the trend components in X_{ij} are removed. This de-trended model residual, as opposed to the raw residual, can better reflect temporal changes in the raw residual. It is possible that other time series model should be fitted when analyzing data of different diseases or from different areas, but the idea summarized in this section is generic. Given that we can estimate μ_i , ϕ_{i1} , ϕ_{i2} and σ_i^2 from (7), the model residual is

$$e_{ij} = X_{ij} - \hat{\mu}_i - \hat{\phi}_{i1}X_{i,j-1} - \hat{\phi}_{i2}X_{i,j-2}.$$

We then have two types of residuals: original/raw residuals, X_{ij} , and model based residuals, e_{ij} , for location s_i on day t_j . Both residuals are useful for detection, because either a statistically significant large daily residual or a large temporal change in the residuals can indicate a possible disease outbreak at a specific time and county. We conduct inference based on the standardized residuals, which are compared on the same scale, as well as the unstandardized residuals which are less variable. In the NC DETECT data, we frequently encounter higher variability in incidence rates for locations with a larger population base. The unstandardized residuals are more sensitive in these scenarios, but the tests based on standardized residuals can have lower power to detect outbreaks.

We use a resampling approach to determine the threshold for significance. This permits detecting abnormalities without being overly sensitive. We outline the procedures below:

1. From the time series model in (7), we estimate $\hat{\mu}_i$, $\hat{\phi}_{i1}$, $\hat{\phi}_{i2}$ and $\hat{\sigma}_i^2$.
2. Critical values for e_{ij} are then obtained by

- Generating replicates $Z_i \sim MVN(0, \widehat{\Sigma}_{S \times S})$, $i = 1, \dots, m$, where $\widehat{\Sigma}_{S \times S} = \text{diag}(\widehat{\sigma}_i^2)$.
 - For each vector Z_i , taking the maximum $\gamma_i = \max_j |Z_{i,j}|$ and $\gamma'_i = \max_j |Z_{i,j}| / \sigma_i$. Let $\Gamma = (\gamma_1, \dots, \gamma_m)$ and $\Gamma' = (\gamma'_1, \dots, \gamma'_m)$. This step is mainly done to control the type I error rate. We obtain the critical value for the unstandardized model filter residuals, obtaining $\Gamma_{(0.95m)}$, and for the standardized ones, obtaining $\Gamma'_{(0.95m)}$, which are the 95% quantiles of Γ and Γ' .
3. Critical values for X_i s are obtained as follows. To generate raw residual sequences, we need to retain the first and the second true X s as baselines. The steps are
- Generate $Z_{i3} \sim N(0, \widehat{\sigma}_i^2)$, $i = 1, \dots, S$.
 - Calculate $X_{i3} = \widehat{\mu}_i + \widehat{\phi}_{i1} X_{i2} + \widehat{\phi}_{i2} X_{i1} + Z_{i3}$, $i = 1, \dots, S$.
 - Among all the generated X_{i3} , take $\delta_1 = \max_i |X_{i3}|$ for the unstandardized and $\delta'_1 = \max_i |X_{i3}| / \sigma_i$ for the standardized versions.
 - Repeat the previous steps m times, and record δ_i and δ'_i each time. Let $\Delta = (\delta_1, \dots, \delta_m)$ and $\Delta' = (\delta'_1, \dots, \delta'_m)$. We obtain the critical value for the unstandardized raw residuals, denoted $\Delta_{(0.95m)}$, and for the standardized ones, denoted $\Delta'_{(0.95m)}$.

Assuming the critical value for raw residuals is c_1 , and the critical value for model residuals is c_2 (the model filtered critical value), we have

$$2 \left(\sqrt{\frac{N(s_i, t_j)}{n(s_i)}} - \sqrt{\widehat{\lambda}(s_i, t_j)} \right) \leq c_1,$$

which leads to

$$\frac{N(s_i, t_j)}{n(s_i)} \leq \left(c_1/2 + \sqrt{\widehat{\lambda}(s_i, t_j)} \right)^2. \quad (8)$$

and, for the model filtered critical value, we have

$$\frac{N(s_i, t_j)}{n(s_i)} \leq \left(\frac{\widehat{\mu}_i + \widehat{\phi}_{i1} X_{i,j-1} + \widehat{\phi}_{i2} X_{i,j-2} + c_2}{2} + \sqrt{\widehat{\lambda}(s_i, t_j)} \right)^2. \quad (9)$$

While using the critical value of the standardized version—denoted c_3 for the scaled raw residuals and c_4 for the scaled model filter residuals—we replace c_1 with $c_3 \widehat{\sigma}_i$ in (8) and c_2 with $c_4 \widehat{\sigma}_i$ in (9). This is very informative in the sense that users can have direct impressions about how large the disease case counts are for determining emerging outbreaks.

3. Data Application and Simulation Study

3.1 Data Application: ED Asthma Admission

The data of interest consist of daily counts of asthma ED admissions from year 2006-2008. For every patient, data on age, sex, county of residence, date of admission to the emergency department, and diagnosis at discharge (coded according to the 9th Revision of the International Classification of Diseases Clinical Modification) are recorded. Mid-year population estimates for the study period were derived using the Population Division of the United States Bureau of the Census county population estimates for NC.

The surveillance system we propose requires that we develop a predictive model for the outcome of interest. Prediction is carried out using existing data for prior days of observation. In our example, we predict daily asthma incidence rates for each county in NC using the prior 2 years of NC DETECT data; as the system is currently configured, we update this predictive model every 90 days in order to incorporate more recent data. At each update, county-specific critical values are obtained and tabled for use in determining whether daily counts are excessive over the next 90 day period.

For the prediction, we need to obtain the population density adjusted distance measure for NC as described in Section 2.2.3. We want to make the local window narrow in order to obtain stable smoothed distances adjusted for the population density, so the bandwidth for the kernel in (6) was determined as follows. Start with $h = 100$, we calculate the population density adjusted distances between counties, $\rho_p(s_i, s_j)|_{h=100}$. Similarly for $h = 50$, we then obtain a discrepancy measure as the summation of differences between distances calculated

when $h = 100$ and $h = 50$, i.e., $\sum_{i,j} \rho_p(s_i, s_j)|_{h=100} - \sum_{i,j} \rho_p(s_i, s_j)|_{h=50}$. Continue the procedure by letting $h = 10, 1$ and 0.1 . When the discrepancy measure stops changing, we consider the smoothed distances to be stabilized, resulting in a bandwidth of 0.1 mile. Figure ?? illustrates Euclidean distance versus population density adjusted distance for the state of North Carolina, and shows the implied North Carolina county layout based on our proposed distance.

The baseline weekly incidences are obtained following the local spatio-temporal approaches discussed in Section 2.2.1. When doing daily incidence estimation, we let cross validation automatically choose the bandwidth a_n in (3) from 0.241, 0.361, 0.482, 0.602 and 0.723, corresponding to $\alpha(14)$, $\alpha(21)$, $\alpha(28)$, $\alpha(35)$ and $\alpha(42)$. In other words, we vary the temporal window from 2 weeks to 6 weeks, and cross validation chooses the best fit. Figure ?? shows the baseline weekly ED asthma admission rate for the 9th, 22nd, 35th and 48th weeks of the year for all the counties of North Carolina. These weeks are shown mainly because they announce the beginning of spring, summer, autumn and winter. Our prediction reflects the seasonal variation in asthma. As can be seen from the figure, the whole state experiences higher incidence in asthma as spring approaches, yet the rates are reduced in the summer season and increase when fall comes. This is expected, considering that allergies, particularly prevalent in the spring and fall due to high pollen levels in the air, can have a huge impact.

The average multiplicative day-of-week effects are: 1.66 on Sunday, 1.44 on Monday, 1.29 on Tuesday, 1.24 on Wednesday, 1.12 on Thursday, 1.16 on Friday and 1.33 on Saturday. With respect to admissions to the ED, it indicates a generally higher rate on Sunday and then followed by Monday, although this effect varies for different counties.

Following the strategies described in Section 2.3, we carry out the residual analysis to identify aberrations from regular patterns. All the counties identified to have potential

outbreaks during the two-year period are shown in Figure ??, which can give users a direct view of these counties' distribution. The map suggests that patterns in the years 2006 and 2007 are very similar: outbreaks are more likely emerging at the northeastern and western boundaries of North Carolina. Because of the loss of neighboring states' information in the analyses, the observed cluster patterns may be explained from the boundary effects, in the sense that people from neighboring counties outside NC may cross the state seeking for health care. This may be also due to regional differences in pollution levels, weather trends and population demographics. It could be due to variations in access to health care outside the ED in these regions of the state as well. Lack of access to good primary care may mean poor management of the disease, or more severe exacerbations, and lack of options for seeking care in a crisis.

We also apply the method prospectively to 2008 data, using a quarterly shifted prediction from historical data. 16 outbreaks were identified, including one outbreak in Hyde from July 22nd to 26th and one in Chowan from October 4th to 8th. There was a large wildfire, which lighting sparked June 1st, in the Pocosin Lakes National Wildlife Refuge, far eastern North Carolina. Asthma chief complaints were significantly high on May 30th in Pamlico and Jones during the pre-fire period. Another alarm was triggered in Alleghany on June 9th, the northwest of North Carolina, probably because the smoke drifted west. Scan statistics identified a primary cluster covering around 30 counties, with Johnston as the center from September 7th to 9th. There is a secondary cluster detected surrounding Ashe from February 9th to 11th. Using the CUSUM method, however, we detected 503 alarms in total. At least one county was flagged every day throughout the year.

3.2 Simulation Study

We create simulated data sets by artificially producing outbreaks in the current data. We want to see if our method can detect emerging outbreaks rapidly and accurately, comparing it to CUSUM and scan statistics.

We chose two counties: Wake, with population 832,875, and Hyde, with population 5449, for adding counts; they represent large and small counties with no known outbreaks in 2006-2007. One simulated outbreak period starts on November 10th, 2006, and the other starts on April 10th, 2007. The prediction of counts for Hyde County during the simulated outbreak periods ranges from 0.03 to 0.49, and for Wake County ranges from 10.56 to 16.70. Generally, we obtain higher admission counts estimates in the spring than in the fall.

We incorporate two types of outbreaks for the simulation: one with a high constant Poisson intensity, and the other with a peak intensity in the middle of the period, with intensity increasing beforehand and decreasing afterward. We consider both short (3 days) and long (7 days) lasting outbreaks to compare different surveillance sensitivities. The counts in the simulated outbreak with constant intensity follow a Poisson distribution with rate λ , $P(\lambda)$. To simulate patterns with a peak, we set the starting date and have the counts initially follow a $P(\lambda_0)$ distribution. During the following days, we first increase and then decrease the intensity by λ_1 every day. Therefore, for example, if we want to generate an outbreak with a peak seven days long, the corresponding intensities for the counts generated are λ_0 , $\lambda_0 + \lambda_1$, $\lambda_0 + 2\lambda_1$, $\lambda_0 + 3\lambda_1$, $\lambda_0 + 2\lambda_1$, $\lambda_0 + \lambda_1$, and λ_0 . We first present the proportions of simulated outbreaks being detected, defined as any day in the simulated period having an alarm, out of 500 simulation runs.

As illustrated in Table ?? and Table ??, we see that CUSUM has extremely high sensitivity at the expense of sounding excessive false alarms. The proposed method actually drastically reduces the number of false alarms as shown in Table ?. While flagging possible aberrations is a priority for a surveillance system, we are hesitant to apply the CUSUM

algorithm due to its oversensitivity from so many detections. The scan statistics approach, on the other hand, is not sensitive especially for small counties. It fails to detect any aberration for Hyde county when the simulated intensity is constantly high during that period. Some other clusters identified include more than 10 counties, which may be too broad for the public health system to take prompt appropriate action.

Therefore, we found the scan statistic to be easily implemented, but somewhat insensitive in detecting outbreaks for small counties, the CUSUM method to have strong sensitivity at the expense of sounding excessive alarms (on average two thousand false alarms in a 2-year period in NC). In contrast, the proposed approach is a reasonable algorithm for real-time outbreak detection, with acceptable sensitivity and a greatly reduced false positive rate.

4. Future Research

The local spatiotemporal estimation method is here applied to aggregated data over counties, i.e., the number of cases in the counties that reported in local EDs. The proposed approach can be naturally extended to point process data. By modeling the space-time intensity of incident cases from accumulated historical data, we can predict the regular pattern if there are no outbreaks. Inference can further be done based on the spatiotemporal residuals obtained after removing the normal trends.

It is important not only to detect potential disease outbreaks but also to evaluate them. When multiple outbreaks are observed, we can next investigate the risk factors associated with such outbreaks. As we mentioned, some demographic characteristics of patients come with surveillance data. Other risk factors of interest include sociodemographic variables, geographic information, and environmental influences. Using normal periods as the control, we can assess the significance of each potential risk factor.

In disease surveillance, there are often many different data streams or different outcomes that we want to monitor. Kulldorff et al. (2007) propose an extension of the spatial and space-time scan statistic that can simultaneously handle multiple data sets. We want to develop techniques to monitor many potential health outcomes simultaneously, such as obesity, asthma, stroke, and unintentional injury, based on our existing work. This can be very challenging, given that we will encounter a fairly complicated situation of having correlated health outcomes in addition to the existing complex spatiotemporal dependence structure.

The new surveillance method will be further developed and validated. It is expected that the new methodology will be generally useful for conducting state-wide or nation-wide public health surveillance.

Acknowledgments

This research was funded in part by a Gillings Innovation Laboratory grant from the University of North Carolina Gillings School of Global Public Health, and Grant no. P30 ES10126 from the Center for Environmental Health & Susceptibility at University of North Carolina at Chapel Hill. NC DETECT data were made available through the NC Public Health Data Group and NC DETECT. The NC Public Health Data Group and NC DETECT do not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

References

- Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. of Royal Statistical Society Series B*. 1974; 36:192–236.

- Craigmile PF, Kim N, Fernandez SA, Bonsu BK. Modeling and detection of respiratory-related outbreak signatures. *BMC Medical Informatics and Decision Making*. 2007; 7
- Cressie, N. *Statistics for Spatial Data*. second edition. Wiley; New York: 1993.
- Cucala, L.; Demattei, C.; Lopes, P.; Ribeiro, A. Spatial scan statistics for case event data based on connected components.
- Daley, D.; Vere-Jones, D. *An Introduction to the Theory of Point Processes*. second edition. Springer; New York: 2003.
- Demattei C, Cucala L. Multiple spatio-temporal cluster detection for case event data : an ordering-based approach. *Communications in Statistics - Theory and methods*. 2011; 40:358–372.
- Diggle PJ, Kaimi I, Abellana R. A kernel method for smoothing point process data. *Applied Statistics*. 1985; 34:138–147.
- Diggle PJ, Rowlingson B, Li Su T. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*. 2005; 16:423–434.
- Fan, J.; Gijbels, I. *Local Polynomial Modelling and its Applications*. Chapman & Hall; London: 1996.
- Hutwagner L, Maloney E, Bean N, L LS, Martin S. Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks. *Emerging Infectious Diseases*. 1997; 3:395–400. [PubMed: 9284390]
- Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. *Statistics in Medicine*. 2007; 26:1594–1507. [PubMed: 16795130]
- Kulldorff M. A spatial scan statistic. *Communications in Statistics - Theory and Methods*. 1997; 26:1481–1496.
- Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *J. of Royal Statistical Society Series A*. 2001; 164:61–72.
- Kulldorff M, Athas W, Feurer E, Miller B, Key C. Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico. *American Journal of Public Health*. 1998; 88:1377–1380. [PubMed: 9736881]
- Kulldorff M, Huang L, Konty K. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*. 2009; 8:58. [PubMed: 19843331]
- Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. *Statistics in Medicine*. 2007; 26:1824–1833. [PubMed: 17216592]
- Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. *Statistics in Medicine*. 1995; 14:799–810. [PubMed: 7644860]
- Møller, J.; Waagepetersen, RP. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC; London: 2004.
- Neill, DB.; Moore, A.; Sabhnani, M.; Daniel, K. Detection of emerging space-time clusters.. *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*; 2005.
- Openshaw S, Charlton M, Wymer C, Craft A. A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*. 1987; 1:335–358.
- Reis B, Mandl K. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*. 2003; 3
- Ripley BD. Modelling spatial patterns (with discussion). *J. of Royal Statistical Society Series B*. 1977; 39:172–212.
- Whittle P. On stationary processes in the plane. *Biometrika*. 1954; 41:434–449.

Figure 1.

The shading shows how use of the proposed population distance metric affects the shape of North Carolina. In particular, the densely-populated area in the center of the state are expanded, while the sparsely populated counties in Western North Carolina are shrunk. In this manner, a similar degree of person-level smoothing is applied regardless of population density.

Figure 2.
Baseline Weekly Asthma Rate (per 100,000 population) of the 9th, 22nd, 35th and 48th Week
in a Year

Figure 3.
Counties in North Carolina Detected for Potential Outbreaks

Table 1

Proportion of Simulated Outbreaks being detected: Constant Pattern

Three Days Outbreak with Constant λ s						
Hyde County	λ					
		1	2	4	6	
	Ratio to Baseline	16.7	33.4	66.7	100.0	
Fall,2006	Proposed Methods	65%	92%	100%	100%	
	CUSUM	95%	100%	100%	100%	
	Scan Statistics	0%	0%	0%	0%	
Spring,2007	Proposed Methods	8%	33%	95%	100%	
	CUSUM	97%	98%	100%	100%	
	Scan Statistics	0%	0%	0%	0%	
Wake County						
	λ	30	40	60	80	
	Ratio to Baseline	2.2	2.9	4.3	5.9	
Fall,2006	Proposed Methods	0%	11%	100%	100%	
	CUSUM	99%	100%	100%	100%	
	Scan Statistics	0.4%	40.2%	50.8%	57.6%	
Spring,2007	Proposed Methods	10%	81%	100%	100%	
	CUSUM	100%	100%	100%	100%	
	Scan Statistics	0%	33.4%	49.2%	42.4%	
Seven Days Outbreak with Constant λ s						
Hyde County	λ					
		1	2	4	6	
	Ratio to Baseline	12.2	24.4	48.9	73.3	
Fall,2006	Proposed Methods	77 %	96%	100%	100%	
	CUSUM	100%	100%	100%	100%	
	Scan Statistics	0%	0%	0%	0%	
Spring,2007	Proposed Methods	14%	60%	99%	100%	

	CUSUM	100%	100%	100%	100%	100%
	Scan Statistics	0%	0%	0%	0%	0%
<hr/>						
Wake County	λ	30	40	60	80	
	Ratio to Baseline	2.1	2.8	4.2	5.6	
<hr/>						
Fall,2006	Proposed Methods	0%	13%	100%	100%	100%
	CUSUM	97%	100%	100%	100%	100%
	Scan Statistics	0%	21.2%	52.6%	52.8%	
<hr/>						
Spring,2007	Proposed Methods	21%	95%	100%	100%	100%
	CUSUM	100%	100%	100%	100%	100%
	Scan Statistics	0%	16.8%	47.4%	47.2%	

The start date of outbreaks for fall, 2006 is November 10th, and the start date of outbreaks for spring, 2007 is April 10th. Each outbreak lasts for 3 (7) days. Simulated counts in Hyde (Wake) County follow Poisson distribution with λ varying as shown in the table.

Table 2

Proportion of Simulated Outbreaks being detected: Peak Pattern

Three Days Outbreak with a Peak						
	λ_1	1	2	4	5	
Hyde County	Mean Ratio to Baseline	31.7	36.2	45.3	49.8	
	Peak Ratio to Baseline	40.7	54.4	81.5	95.1	
Fall,2006	Proposed Methods	100%	100%	100%	100%	
	CUSUM	100%	100%	100%	100%	
	Scan Statistics	0.4%	17.2%	11.2%	5.8%	
Spring,2007	Proposed Methods	78%	86%	97%	99%	
	CUSUM	100%	100%	100%	100%	
	Scan Statistics	0%	17.2%	11.2%	5.6%	
λ_1						
		5	10	20	25	
Wake County	Ratio to Baseline	2.1	2.2	2.5	2.6	
	Peak Ratio to Baseline	2.3	2.7	3.5	4.9	
Fall,2006	Proposed Methods	1%	6%	59%	82%	
	CUSUM	94%	100%	100%	100%	
	Scan Statistics	5.4%	52.2%	49.2%	50%	
Spring,2007	Proposed Methods	23%	29%	92%	98%	
	CUSUM	93%	100%	100%	100%	
	Scan Statistics	5%	46.2%	50.2%	50%	
Seven Days Outbreak with a Peak						
	λ_1	1	2	4	5	
Hyde County	Mean Ratio to Baseline	7.2	10.0	15.6	18.4	
	Peak Ratio to Baseline	10.9	17.5	30.6	37.1	
Fall,2006	Proposed Methods	99%	100%	100%	100%	
	CUSUM	100%	100%	100%	100%	

Spring,2007	Scan Statistics	32.4%	9.4 %	0%	0%
	Proposed Methods	88%	99%	100%	100%
	CUSUM	100%	100%	100%	100%
	Scan Statistics	32.6%	6.6%	0%	0%
<hr/>					
Wake County	λ_1	5	10	15	25
	Mean Ratio to Baseline	2.1	2.5	3.3	3.8
	Peak Ratio to Baseline	2.6	3.6	5.6	6.6
	<hr/>				
Fall,2006	Proposed Methods	1%	51%	100%	100%
	CUSUM	83%	99%	100%	100%
	Scan Statistics	53.2%	51.8%	49%	49.4%
	<hr/>				
Spring,2007	Proposed Methods	70%	100%	100%	100%
	CUSUM	81%	99%	100%	100%
	Scan Statistics	46.8%	48.2%	51%	50.6%
	<hr/>				

The start date of outbreaks for fall, 2006 is November 10th, and the start date of outbreaks for spring, 2007 is April 10th. Each outbreak lasts for 3 (7) days. Simulated counts in Hyde (Wake) follow Poisson distribution with starting/ending days intensity $\lambda_0 = 2$ ($\lambda_0 = 25$) and intensity difference between consecutive days λ_1 as shown in the table.

Table 3
Number of False Outbreaks Detected, Seven-Day Peak Pattern, Year 2006-2007

λ_1 (Hyde)	1	2	3	4	5
λ_1 (Wake)	5	10	15	20	25
Number of Simulated Outbreaks	28	28	28	28	28
Maximum Number of Outbreaks	73000	73000	73000	73000	73000
Number of False Outbreaks	Proposed Method				
	137	136	134	128	127
Number of False Outbreaks	CUSUM				
	2889	2889	2889	2889	2889