Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

International Journal of Forecasting 25 (2009) 498–517

international journal of forecasting

www.elsevier.com/locate/ijforecast

Expectation-based scan statistics for monitoring spatial time series data

Daniel B. Neill*

H.J. Heinz III College, School of Public Policy and Management, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

Abstract

We consider the simultaneous monitoring of a large number of spatially localized time series in order to detect emerging spatial patterns. For example, in disease surveillance, we detect emerging outbreaks by monitoring electronically available public health data, e.g. aggregate daily counts of Emergency Department visits. We propose a two-step approach based on the *expectation-based scan statistic*: we first compute the expected count for each recent day for each spatial location, then find spatial regions (groups of nearby locations) where the recent counts are significantly higher than expected. By aggregating information across multiple time series rather than monitoring each series separately, we can improve the timeliness, accuracy, and spatial resolution of detection. We evaluate several variants of the expectation-based scan statistic on the disease surveillance task (using synthetic outbreaks injected into real-world hospital Emergency Department data), and draw conclusions about which models and methods are most appropriate for which surveillance tasks.

Keywords: Time series monitoring; Pattern detection; Event detection; Spatial scan statistics; Biosurveillance

1. Introduction

Many applications require the monitoring of time series data in order to detect anomalous counts. A traditional application of time series monitoring is the use of statistical process control to ensure consistency in manufacturing: the process is measured regularly to ensure that the desired specifications (e.g. product size and weight) remain within an acceptable range. More recently, time series monitoring has been used in a variety of event detection systems: crime surveillance systems (Gorr & Harries, 2003; Levine, 1999) detect emerging hot-spots of crime activity, disease surveillance systems (Sabhnani et al., 2005) monitor electronic public health data such as hospital visits and medication sales in order to detect emerging outbreaks, and environmental monitoring systems (Ailamaki, Faloutsos, Fischbeck, Small, & VanBriesen, 2003) detect abnormally high pollutant levels in the air, water, and soil.

* Tel.: +1 412 268 3885.
E-mail address: neill@cs.cmu.edu.

0169-2070/$ - see front matter © 2008 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2008.12.002

In all of these event detection applications, we wish to detect emerging spatial patterns as quickly and accurately as possible, enabling a timely and appropriate response to the detected events. As a concrete example, we focus on the task of detecting outbreaks of respiratory illness using hospital Emergency Department (ED) data. In this case, we can monitor the number of patients visiting the ED with respiratory symptoms in each zip code on each day. Each zip code s_i has a corresponding time series of daily counts c_i^t, and our goal is to detect anomalous increases in counts that correspond to an emerging outbreak of disease.

A variety of methods have been developed to monitor time series data and detect emerging anomalies. Control chart methods (Shewhart, 1931) compare each observed count to its expected value (a counterfactual forecast obtained from time series analysis of the historical data), and detect any observations outside a critical range. Cumulative sum methods (Page, 1954) and tracking signals (Brown, 1959; Trigg, 1964) aggregate these deviations across multiple time steps in order to detect shifts in a process mean. When extending these techniques to the simultaneous monitoring of multiple time series, we have several options (Burkom, Murphy, Coberly, & Hurt-Mullen, 2005). In the simplest, "parallel monitoring" approach, we monitor each time series separately and report any anomalous values. In the "consensus monitoring" approach, we combine the signals from multiple time series in order to achieve higher detection power. To detect anomalies that affect multiple time series simultaneously, we can either combine the outputs of multiple univariate detectors or treat the multiple time series as a single multivariate quantity to be monitored. For example, multivariate control charts (Hotelling, 1947) learn the joint distribution of a set of signals from historical data, and detect when the current multivariate signal is sufficiently far from its expectation.

We note, however, that none of these time series monitoring methods account for the *spatial* nature of the event detection problem. We expect events to be localized in space: if a given location is affected by the event, nearby locations are more likely to be affected than locations that are spatially distant. For example, disease outbreaks tend to affect spatially contiguous areas, either because of contagion (e.g. human-to-human transmission) or because the cases share a common source (e.g. contaminated drinking water). Thus, we must consider alternate methods of monitoring spatial time series data, where we expect anomalies to affect the time series for some spatially localized subset of locations.

A typical approach to the monitoring of spatial time series data uses "fixed partitions": we map the locations to a Euclidean space (e.g. using the longitude and latitude of each zip code centroid), partition the search space such that each location is contained in exactly one partition, and aggregate the counts for each partition into a single time series. We then monitor the time series for each partition separately, and report any anomalous counts. One challenge is deciding how to partition the search space: in the case of zip code level data, we could consider each zip code to be a separate partition, combine multiple adjacent zip codes in a single partition, or even aggregate all of the zip codes into a single time series. An alternative is to form an "ad-hoc partitioning" by identifying individual locations with high counts and using some heuristic to cluster these locations (Corcoran, Wilson, & Ware, 2003).

Any choice of partitioning scheme creates a set of potential problems, which we call the "curse of fixed partitions". In general, we do not have *a priori* knowledge of how many locations will be affected by an event, and we wish to maintain high detection power whether the event affects a single location, all locations, or anything in between. A coarse partitioning of the search space will lose power to detect events that affect a small number of locations, since the anomalous time series will be aggregated with other counts that are not anomalous. A fine partitioning of the search space will lose power to detect events that affect many locations, since only a small number of anomalous time series are considered in each partition. Partitions of intermediate size will lose some power to detect both very small and very large events. Moreover, even if the partition size corresponds well to the event size, the fixed partition approach will lose power if the affected set of locations is divided between multiple partitions rather than corresponding to a single partition. While ad-hoc partitioning methods allow partitions to vary in size, the chosen set of partitions still may not correspond to the actual set of locations affected, resulting in a loss of detection power.

Our solution to the "curse of fixed partitions" is a multi-resolution approach in which we search over a large and overlapping set of spatial regions, each containing some subset of the spatial locations, and find the most significant clusters of anomalous counts. Because we search over both large regions (coarse resolutions) and small regions (fine resolutions), this method has high power to detect both large and small clusters. Similarly, by searching over regions with varying shape, size, and duration, we can achieve high detection power for clusters with a wide range of spatial and temporal characteristics.

More precisely, we propose an "expectation-based scan statistic" approach with two distinct steps. First, we compute the "expected count" for each spatial location for each recent day of data. Each expected count is a counterfactual forecast assuming that no clusters are present, and is obtained from time series analysis of the historical counts for that spatial location. The second step is to detect space–time clusters, i.e. groups of nearby locations where the recent counts are significantly higher than expected. To do so, we develop a new variant of the spatial and space–time scan statistics (Kulldorff, 1997, 2001; Neill & Moore, 2005) which searches over space–time regions, compute a likelihood ratio statistic for each region, and detect the most significant clusters.

In the remainder of this paper, we describe the expectation-based scan statistic, and evaluate the performance of this method on the disease surveillance task. In Section 2, we present an overview of the expectation-based scan approach, and in Sections 3 and 4 we discuss each of the two steps (computing expected counts, and finding clusters with higher than expected counts) in detail. Finally, Sections 5 and 6 present and discuss our set of evaluations using hospital Emergency Department data, comparing several variants of the expectation-based scan and demonstrating large performance improvements as compared to traditional fixed partition and spatial scan approaches.

2. The expectation-based scan statistic

As noted above, the event detection problem requires us to monitor a set of spatially localized time series in order to detect emerging spatial clusters of anomalous counts. We assume a given set of spatial locations $\{s_i\}$, where each location s_i corresponds to a point in Euclidean space. For example, in our disease surveillance task, each zip code is mapped to a two-dimensional space using the longitude and latitude of the zip code centroid. For each location s_i, we are given a time series of non-negative integer counts c_i^t, where $t = 0$ represents the current time step and $t = 1 \ldots t_{\max}$ represent the historical data from 1 to $t_{\max}$ time steps ago respectively. For example, each count c_i^t could represent the number of respiratory Emergency Department visits in zip code s_i on day t.

Given this data, our goal is to detect any spatial region where the recent counts are significantly higher than expected. More precisely, we define a set of "space–time regions" $\mathbf{S}$, where each region $S \in \mathbf{S}$ contains a subset of spatial locations $\{s_i : s_i \in S\}$, and also has a time duration $w(S)$, meaning that the given set of spatial locations has been affected for the most recent w time steps ($t = 0 \ldots w - 1$). We search over space–time regions with durations $w = 1 \ldots W$, where W is the "maximum temporal window size"; i.e., we are only interested in clusters that have emerged within the past W time steps.

The expectation-based scan statistic has two distinct steps: we first compute the "expected count" (or "baseline") b_i^t for each spatial location s_i and each time step $t = 0 \ldots W - 1$, and then detect space–time regions S with higher than expected counts. For the first step, each baseline b_i^t is a counterfactual estimate of c_i^t assuming the null hypothesis of no clusters, and is computed from the historical data in location s_i using some method of time series analysis. We must then find space–time regions S where the observed counts c_i^t are significantly higher than the expected counts b_i^t, for locations $s_i \in S$ and time steps $t = 0 \ldots w(S) - 1$.

To do so, we develop a likelihood ratio test based on the spatial and space–time scan statistics. The spatial scan statistic, first presented by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) and extended to space–time data by Kulldorff, Athas, Feuer, Miller, and Key (1998) and Kulldorff (2001), is commonly used in the public health community for purposes ranging from the detection of bioterrorist attacks (Neill, 2006) and emerging infectious diseases (Mostashari, Kulldorff, Hartman, Miller, & Kulasekera, 2003) to the identification of environmental risk factors for breast cancer (Kulldorff, Feuer, Miller, & Freedman, 1997) and childhood leukemia (Hjalmars, Kulldorff, Gustafsson, & Nagarwalla, 1996). Here we apply the generalized spatial scan framework described by Neill and Moore (2005). We first define generative models of the data under H_0, the null hypothesis of no clusters, and under $H_1(S)$, the alternative hypothesis assuming a cluster in some space–time region S. We then compute the likelihood ratio statistic $F(S)$ for each space–time region $S \in \mathbf{S}$. The likelihood ratio is defined as the ratio of the data likelihoods under the alternative and null hypotheses:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}.$$

If the null or alternative hypotheses have any free parameters, we compute the likelihood ratio using the maximum likelihood estimates of each parameter:

$$F(S) = \frac{\max\limits_{\theta_1(S)\in\Theta_1(S)} \Pr(\text{Data} \mid H_1(S), \theta_1(S))}{\max\limits_{\theta_0\in\Theta_0} \Pr(\text{Data} \mid H_0, \theta_0)}.$$

The regions with the highest values of the likelihood ratio statistic are those which are most likely to have been generated under the alternative hypothesis (cluster in space–time region S) instead of the null hypothesis of no clusters. However, because we are maximizing the likelihood ratio over a large number of regions, we are very likely to find many regions with high likelihood ratios even when the null hypothesis is true. Thus we must choose a threshold value F_{thresh} based on the acceptable false positive rate, and report all space–time regions S with $F(S) > F_{\text{thresh}}$.

The detection threshold F_{thresh} can be calibrated in one of two ways. The traditional approach (Kulldorff, 1997) is to perform randomization testing: we generate a large number of replica datasets under the null hypothesis and compute the distribution of the maximum region score under the null. Then, for a given false positive rate α, F_{thresh} is the $100(1-\alpha)$ percentile value of the null distribution. An alternative is to use the empirical distribution of maximum region scores from a large amount of historical data, again using the $100(1-\alpha)$ percentile value from this distribution. Neill (2007) compared these two approaches on real-world Emergency Department and over-the-counter medication sales data, and concluded that randomization testing tends to be oversensitive (producing false positive rates of up to 40% for $\alpha = 0.05$), and that the empirical approach tends to achieve higher detection power. Nevertheless, randomization testing may be a useful alternative to the empirical approach when only a limited amount of historical data is available.

3. Time series forecasting

A variety of time series analysis methods may be used to compute counterfactual forecasts b_i^t for each spatial location s_i, using the time series of historical counts c_i^t. We compute these baselines (expected counts) for each of the past W days, where W is the maximum temporal window size, then search over space–time regions with time durations up to W. Here we consider eight different time series analysis methods: four simple "moving average" methods, two "day-of-week adjusted moving average" methods, and two methods that account for both day-of-week and seasonality. Letting $t = 0$ represent the current day's counts, and $t > 0$ represent the historical counts from t days ago, the 28-day moving average (MA28) is defined as follows:

$$b_i^t = \frac{1}{28} \sum_{u=t+1\ldots t+28} c_i^u.$$

The 7-day, 14-day, and 56-day moving averages are defined analogously, and we compare all four of these methods in our evaluation below. While such methods may be sufficient for datasets without seasonal or day-of-week trends, datasets with these trends may require more complex forecasting methods. The simple moving average may be adjusted for day of the week by estimating the proportion of counts β_i^j occurring in each location s_i on each day of the week ($j = 1\ldots7$). Then when we predict the baseline value for a given location on a given day, we choose the corresponding value of β_i^j and multiply our estimate by $7\beta_i^j$. We distinguish between "local" and "global" methods of adjusting for day of the week, where the local method computes β_i^j using only counts from location s_i, and the global method computes β_i^j using the global aggregate counts $g_t = \sum_i c_i^t$. Using 12 weeks of historical data, we can compute the β_i^j for the local (MALD) and global (MAGD) methods respectively:

$$\text{MALD:}\quad \beta_i^j = \frac{\sum_{t=j,j+7,\ldots,j+77} c_i^t}{\sum_{t=1,\ldots,84} c_i^t},$$

$$\text{MAGD:}\quad \beta_i^j = \frac{\sum_{t=j,j+7,\ldots,j+77} g_t}{\sum_{t=1,\ldots,84} g_t}.$$

The global adjustment for day of the week assumes that weekly trends have a constant and multiplicative effect on counts for each spatial location. The local adjustment assumes that each spatial location displays different weekly trends, and thus accounts for space by day-of-week interaction. On the other hand, we expect the global estimates to display less variance since they rely on larger counts to estimate the day-of-week proportions.

Finally, we consider two other methods of time series forecasting, both of which account not only for day-of-week trends but for seasonal trends as well. The multiplicative Holt–Winters' (HW) method is a commonly used extension of exponential smoothing that also adjusts dynamically for cyclical day of the week effects and for linear trends (e.g. those due to seasonal variation in counts); it was shown to be highly effective for temporal biosurveillance by Burkom, Murphy, and Shmueli (2007). The Holt–Winters' forecasts are calculated from the counts c_i^t by iterating the following three equations for the smoothed value S_t, trend component T_t, and day-of-week component I_t, from $t = 84$ to $t = 1$ (the day before the current day $t = 0$):

$$S_t = \alpha \frac{c_i^t}{I_{t+7}} + (1-\alpha)(S_{t+1} + T_{t+1})$$
$$T_t = \beta(S_t - S_{t+1}) + (1-\beta)T_{t+1}$$
$$I_t = \gamma \frac{c_i^t}{S_t} + (1-\gamma)I_{t+7}.$$

Then each day's expected count is given by the one-step-ahead, day-of-week adjusted estimate, $b_i^t = (S_{t+1} + T_{t+1})I_{t+7}$. For our comparison, we used fixed values of $\alpha = \beta = \gamma = 0.1$; these were not optimized, but achieved adequate performance in preliminary studies on a different public health dataset.

Finally, the "current day" (CD) method was used to derive baseline estimates for the space–time permutation statistic (Kulldorff, Heffernan, Hartman, Assuncao, & Mostashari, 2005). This method assumes that counts are independently distributed in space and time, and thus the expected count b_i^t for a given location s_i on a given day t is equal to the total number of cases on day t multiplied by the fraction of cases occurring in location s_i:

$$b_i^t = \frac{\sum_i c_i^t \sum_t c_i^t}{\sum_i \sum_t c_i^t}.$$

This method focuses on detecting space–time interaction, and thus it does not detect purely spatial or purely temporal clusters. As noted by Neill, Moore, Sabhnani, and Daniel (2005b), this method uses the current day's aggregate count to estimate the current day's counts, and may lose power to detect spatially large clusters. On the other hand, by conditioning on the aggregate count, CD can produce accurate baseline estimates even in the presence of strong temporal trends due to day of the week, seasonality, or holidays.

4. Detecting space–time clusters

The primary advantage of the scan statistic approach is that, rather than detecting individual counts c_i^t that are higher than expected, it integrates information over multiple spatial locations s_i and multiple time steps t to detect space–time clusters with higher than expected counts. Unlike typical methods that use a fixed partition of the search space, the level of aggregation need not be determined in advance. Instead, we search over a large set of space–time regions with varying sizes, shapes, and temporal durations. This process can be visualized as moving a "space–time window" around the search area, allowing the size, shape, and duration of this window to vary, and detecting any window which contains anomalously high counts. As noted above, we choose a set of search regions $\mathbf{S}$, where each region $S \in \mathbf{S}$ contains a set of spatial locations $\{s_i : s_i \in S\}$, and also has a temporal duration $w(S)$. For the alternative hypothesis $H_1(S)$, representing a cluster in region S, we assume that the locations $s_i \in S$ have been affected for the most recent $w(S)$ time steps. Under the null hypothesis H_0, no locations are affected. For notational simplicity, we let $c_i^t \in S$ denote the set of counts in region S, i.e. those counts c_i^t such that $s_i \in S$ and $t = 0 \ldots w(S) - 1$. For each region S under consideration, we evaluate the likelihood ratio score $F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$, comparing the counts $c_i^t \in S$ to their expected values b_i^t. We now consider the choice of search regions, and the computation of the likelihood ratio statistic, in more detail.

4.1. Search regions

Our set of search regions $\mathbf{S}$ was determined by mapping the spatial locations (e.g. zip code centroids) to a uniform $N \times N$ grid, and then searching over all rectangular regions on the grid. In this case, each space–time region $S \in \mathbf{S}$ can be defined by five integers $(x_{\min}, y_{\min}, x_{\max}, y_{\max}, w)$, representing the x and y coordinates of the top left and bottom right grid cells, and the time duration, respectively. All combinations of $0 \leq x_{\min} \leq x_{\max} < N$, $0 \leq y_{\min} \leq y_{\max} < N$, and $1 \leq w \leq W$ were considered. Thus, the total number of search regions increases linearly with the maximum window size W, and proportional to the fourth power of the grid size N. For each region S, the score $F(S)$ can be calculated from the counts $c_i^t \in S$ and baselines $b_i^t \in S$. Neill and Moore (2004) and Neill, Moore, and Sabhnani (2005a) demonstrated that these scores can be found in a computationally efficient manner, calculating the aggregate count $\sum c_i^t$ and aggregate baseline $\sum b_i^t$ for each rectangular region S in constant time, and then using these sufficient statistics to calculate the score $F(S)$. For large grid sizes, further computational speedups can be gained by using the *fast spatial scan* (Neill & Moore, 2004; Neill, Moore, Pereira, & Mitchell, 2005) to find the highest scoring clusters without an exhaustive search.

We note that this approach considers a wide range of region sizes, varying from individual 1×1 cells to the entire $N \times N$ grid, and thus has high power to detect both small and large clusters. Similarly, it considers regions ranging from squares to highly elongated rectangles, and thus has high power to detect both compact and elongated clusters. Finally, by allowing the cluster duration (i.e. the window size w) to vary, it has high power to detect both rapidly and gradually emerging clusters.

4.2. The expectation-based Poisson (EBP) statistic

As noted above, we must evaluate the likelihood ratio score $F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$ for each spatial region $S \in \mathbf{S}$, and report the highest scoring regions. To do so, we must choose models of how the data is generated, both under the null hypothesis H_0 (assuming that no clusters are present) and under the set of alternative hypotheses $H_1(S)$, each representing a cluster in some region S.

We assume that each count c_i^t has been drawn from a Poisson distribution. This is a common assumption in the epidemiological literature: individual cases can be specified as the realization of a Poisson point process (with spatially varying intensity based on the population at risk), and thus the aggregate count in each spatial area is Poisson distributed. We allow the Poisson means to vary over both space and time: under the null hypothesis H_0 of no clusters, we assume that each count c_i^t has been drawn with mean equal to the expected count b_i^t. As discussed in Section 3, these baselines are learned from the historical data for location s_i by time series analysis. Under the alternative hypothesis $H_1(S)$, we assume that each count $c_i^t \in S$ has been drawn with mean equal to the product of the expected count b_i^t and some constant q_i^t, which we call its *relative risk*. Thus, we wish to search for regions S where the relative risks q_i^t are greater than 1. We make the further simplifying assumption that the relative risks are uniform over region S ($q_i^t = q$ for all $c_i^t \in S$), and thus we expect a constant multiplicative increase in counts in the affected region.

As noted above, we wish to determine whether any space–time region S has significantly higher than expected counts. Under our model assumptions, where each count $c_i^t \in S$ has been drawn from a Poisson distribution with mean proportional to the baseline b_i^t times the uniform relative risk q, this question simplifies to determining whether any region S has $q > 1$. Thus, we test the null hypothesis H_0 against the set of alternative hypotheses $H_1(S)$, where:

H_0: $c_i^t \sim \text{Poisson}(b_i^t)$ for all spatial locations s_i and time steps t.

$H_1(S)$: $c_i^t \sim \text{Poisson}(q b_i^t)$ for all $c_i^t \in S$, and $c_i^t \sim \text{Poisson}(b_i^t)$ for all $c_i^t \notin S$, for some $q > 1$.

Computing the likelihood ratio, and using the maximum likelihood estimate for the parameter q (the uniform relative risk for region S), we obtain the following expression:

$$F(S) = \frac{\max_{q>1} \prod_{c_i^t \in S} \Pr(c_i^t \sim \text{Poisson}(qb_i^t))}{\prod_{c_i^t \in S} \Pr(c_i^t \sim \text{Poisson}(b_i^t))}.$$

Plugging in the equations for the Poisson likelihood, and simplifying, we obtain:

$$F(S) = \frac{\max_{q>1} \prod_{c_i^t \in S} \mathrm{e}^{-qb_i^t}(qb_i^t)^{c_i^t}/(c_i^t)!}{\prod_{c_i^t \in S} \mathrm{e}^{-b_i^t}(b_i^t)^{c_i^t}/(c_i^t)!}$$

$$= \max_{q>1} \prod_{c_i^t \in S} \mathrm{e}^{(1-q)b_i^t} q^{c_i^t} = \max_{q>1} \mathrm{e}^{(1-q)B} q^C,$$

where C and B are the total count $\sum c_i^t$ and total baseline $\sum b_i^t$ of region S, respectively. We find that the value of q that maximizes the numerator is $q = \max(1, \frac{C}{B})$. Plugging in this value of q, we obtain:

$$F(S) = \left(\frac{C}{B}\right)^C \mathrm{e}^{B-C}$$

if $C > B$, and $F(S) = 1$ otherwise. Because $F(S)$ is a function only of the sufficient statistics C and B, this function is efficiently computable: we can calculate the score of any region S by first calculating the aggregate count and baseline, and then applying the function F.

4.3. Comparison to Kulldorff's statistic

The spatial scan statistic was originally presented by Kulldorff (1997) in a purely spatial setting, where we are given a single count c_i and baseline b_i for each spatial location s_i. Later work (Kulldorff, 2001; Kulldorff et al., 1998) extended this approach to the space–time scan statistic, where we are given the counts c_i^t and baselines b_i^t for each time step t. In these settings, the baselines b_i^t are given, and are assumed to represent the population of each location s_i. Under the null hypothesis of no clusters, we assume a spatially uniform incidence rate $q_i^t = q_{all}$, and thus we expect the counts c_i^t to be proportional to the baselines b_i^t. Under the alternative hypothesis $H_1(S)$, we assume that the incidence rate is higher inside the region than outside. Kulldorff's statistic can also be used in our expectation-based scan statistic framework, where the baselines are expected counts learned from time series analysis of historical data. In this case, the null hypothesis H_0 assumes a constant relative risk q_{all}, and the alternative hypothesis $H_1(S)$ assumes a higher relative risk inside region S than outside ($q_{in} > q_{out}$).

More precisely, Kulldorff's statistic assumes that each count c_i^t is generated independently from a Poisson distribution with mean proportional to the expected count b_i^t times the relative risk q_i^t. Furthermore, we assume that the relative risk is uniform both inside the region ($q_i^t = q_{in}$ for all $c_i^t \in S$) and outside the region ($q_i^t = q_{out}$ for all $c_i^t \notin S$). Then we test the null hypothesis H_0 against the set of alternative hypotheses $H_1(S)$, where:

H_0: $c_i^t \sim \text{Poisson}(q_{all} b_i^t)$ for all locations s_i and time steps t, for some constant q_{all}.

$H_1(S)$: $c_i^t \sim \text{Poisson}(q_{in} b_i^t)$ for all $c_i^t \in S$, and $c_i^t \sim \text{Poisson}(q_{out} b_i^t)$ for all $c_i^t \notin S$, for some constants $q_{in} > q_{out}$.

In this case, the alternative hypothesis has two free parameters (q_{in} and q_{out}) and the null hypothesis has one free parameter (q_{all}). Computing the likelihood ratio, and using the maximum likelihood parameter estimates $q_{in} = \frac{C_{in}}{B_{in}}$, $q_{out} = \frac{C_{out}}{B_{out}}$, and $q_{all} = \frac{C_{all}}{B_{all}}$, we obtain the following expression for the likelihood ratio:

$$F(S) = \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}$$

if $\frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}}$, and $F(S) = 1$ otherwise. In this expression, C_{in} and B_{in} are the total count $\sum c_i^t$ and total baseline $\sum b_i^t$ inside region S, C_{out} and B_{out} are the total count and baseline outside region S, and C_{all} and B_{all} are the total count and baseline everywhere. A detailed derivation of Kulldorff's statistic is provided by Kulldorff (1997) and Neill (2006).

We note that the assumptions made by Kulldorff's statistic are very different from the assumptions made by EBP, resulting in a substantially different likelihood ratio statistic. EBP expects the counts c_i^t to be *equal* to the baselines b_i^t under the null hypothesis, not just *proportional* to the baselines, since in this case the baselines represent expected counts. Under the alternative hypothesis $H_1(S)$, Kulldorff's statistic compares the relative risks inside and outside the region, while EBP ignores the counts outside the region and simply compares the relative risk inside the region to 1. Assuming that we can accurately estimate the expected count (under the null hypothesis of no clusters) in each spatial location, and that we are interested in detecting any regions with higher than expected counts, the EBP statistic is a more natural model representation. Additionally, Kulldorff's statistic will lose power to detect spatially large clusters. Let us consider the extreme case where a cluster causes a uniform multiplicative increase in counts over the entire search region. In this case, we would have $q_{in} = q_{out} = q_{all} \gg 1$: Kulldorff's statistic would be entirely unable to detect this increase, while EBP would easily detect it. On the other hand, Kulldorff's statistic has the advantage of being more robust to misestimation of global trends such as day-of-week and seasonality, since the parameter q_{all} automatically adjusts for the case when all estimates are incorrect by a constant multiplicative factor. Thus, it is an open question as to which methodology will work better in real-world time series monitoring scenarios.

5. Evaluation

To evaluate the detection performance of the expectation-based scan statistic in the disease surveillance domain, we obtained one year of Emergency Department data from Allegheny County, Pennsylvania. Daily counts of the number of patient records with respiratory chief complaints were aggregated at the zip code level, and thus each count c_i^t represented the number of respiratory ED visits for a given zip code s_i for a given day t. The ED dataset contains data for 88 distinct Allegheny County zip codes from January 1, 2002 to December 31, 2002. The first 84 days of data were used for baseline calculations only, giving us 281 days of count data for evaluation. The total number of respiratory ED visits in Allegheny County ranged from 5 to 62 cases per day, with a mean of 32.58 cases and standard deviation of 7.60 cases. To test for day-of-week trends in the data, we performed a one-way ANOVA using the aggregate daily count as the dependent variable and day of the week as the independent variable. This analysis revealed that counts tend to be significantly higher on Mondays than on other days of the week. Similarly, ANOVA analysis using month of the year as the independent variable revealed that counts tend to be significantly higher during the winter months (November through March) and significantly lower during the summer months (May through August). These results suggest the potential value of using more complex time series forecasting methods that can account for day-of-week and seasonal trends, and we consider several such methods in Section 5.3.

To evaluate the detection power of our methods, we used a semi-synthetic testing framework, injecting simulated outbreaks of disease into the real-world data. In the biosurveillance domain, outbreak simulation is commonly used to evaluate detection methods because of the scarcity of available, labeled data from real-world outbreaks. Simulation also allows us to precisely evaluate the effect of different parameters (such as the size, shape, and temporal progression of the outbreak) on each method's detection performance; as we observe below, many of these parameters have substantial effects on the relative performance of different methods. Simulation of outbreaks is an active area of ongoing research in biosurveillance, and several recently developed methods (Buckeridge et al., 2004; Wallstrom, Wagner, & Hogan, 2005) show great promise for producing realistic outbreak scenarios. Finally, we note that the expectation-based scan statistic method has been evaluated retrospectively on the 2000 gastroenteritis outbreak which occurred in Walkerton, Ontario, and was able to detect the outbreak two days before the first public health response (Davies, for the ECADS partners and collaborators, 2006; Neill, 2006). However, data from a single outbreak is insufficient to draw detailed conclusions about the relative performance of different methods.

We considered a simple class of circular outbreaks with a linear increase in the expected number of cases over the duration of the outbreak. More precisely, our outbreak simulator takes four parameters: the outbreak duration T, the outbreak severity Δ, and the minimum and maximum number of zip codes affected, $k_{\min}$ and $k_{\max}$. Then for each injected outbreak, the outbreak simulator chooses the start date of the outbreak t_{start}, the number of zip codes affected k, and the center zip code s_{center} uniformly at random. The outbreak is assumed to affect zip code s_{center} and its $k-1$ nearest neighbors, as measured by the distance between the zip code centroids. On each day t of the outbreak, $t = 1 \ldots T$, the outbreak simulator injects Poisson($tw_i\Delta$) cases into each affected zip code, where w_i is the "weight" of each affected zip code, set proportional to its total count $\sum_t c_i^t$ for the entire dataset, and normalized so that the total weight equals 1 for each injected outbreak. Using this simple outbreak simulator, we performed three simulations of varying size: "small" injects affecting 1–10 zip codes, "medium" injects affecting 10–20 zip codes, and "large" injects affecting all zip codes in Allegheny County. We used $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$ for small, medium, and large injects respectively. We used a value of $T = 7$ for these outbreaks, and thus outbreaks were assumed to be one week in duration. An additional simulation of a "gradual" outbreak affecting 10–20 zip codes, with $T = 28$ and $\Delta = 1$, was used to evaluate scan statistics with varying temporal window sizes, as discussed below. For each of these simulations, we considered 1000 different, randomly generated outbreaks.

For each combination of method and outbreak size, we computed the method's proportion of outbreaks detected, and the average number of days to detection, as a function of the allowable false positive rate. To do this, we first computed the maximum region score $F^* = \max_S F(S)$ for each day of the original dataset with no outbreaks injected (as noted above, the first 84 days of data are excluded, since these are used to calculate baselines for our methods). Then, for each injected outbreak we computed the maximum region score for each outbreak day, and determined the proportion of the days for which the original dataset had higher scores. Assuming that the original dataset contains no outbreaks, this is the proportion of false positives that we would have to accept in order to have detected the outbreak on day t. For a fixed false positive rate r, the "days to detect" for a given outbreak is computed as the first outbreak day ($t = 1 \ldots 7$) with the proportion of false positives less than r. If no day of the outbreak has the proportion of false positives less than r, the method has failed to detect that outbreak: for the purposes of our "days to detect" calculation, these are penalized proportional to the length of the outbreak, and thus are counted as 14 days to detect for a 7-day outbreak. The tradeoff between the false positive rate and average days to detection for a given method can be visualized as an Activity Monitoring Operating Characteristic (AMOC) curve (Fawcett & Provost, 1999), or we can compare detection times for a fixed false positive rate such as 1 false positive per month. In either case, a lower detection time for a given false positive rate corresponds to improved detection performance.

A secondary performance measure is the method's *spatial detection accuracy*, which measures its ability to precisely pinpoint the set of locations affected by an outbreak. For a given day of a simulated outbreak, we define the "true outbreak region" S_{true} as the set of locations with injected cases, and we define the "detected region" S^* as the region with the highest score, $S^* = \arg\max_S F(S)$. Then the *spatial precision* is defined as the ratio of correctly detected locations to all detected locations, $\frac{\#\{s_i \in S^* \cap S_{true}\}}{\#\{s_i \in S^*\}}$. Similarly, the *spatial recall* is defined as the ratio of correctly detected locations to all correct locations, $\frac{\#\{s_i \in S^* \cap S_{true}\}}{\#\{s_i \in S_{true}\}}$. For each simulated outbreak, we compute the precision and recall at the midpoint of the outbreak (e.g. day 4 of a 7-day outbreak), and average the precision and recall measures over all outbreaks for a given simulation. Finally, we compute the *F-measure* (harmonic mean of precision and recall), as an aggregate measure of the method's spatial detection accuracy.

5.1. Comparison of the scan statistic and fixed partition approaches

In our first set of experiments, we compared the scan statistic method to the traditional "fixed partition" method of time series surveillance. We considered a range of grid sizes from $N = 1$ to $N = 32$. For each grid size, we evaluated the performance of the scan statistic (searching over all rectangular regions on the grid, with sizes varying from 1×1 to $N \times N$) and the fixed partition approach (treating each grid cell as a separate time series, and thus searching over 1×1 regions only). As noted above, the mapping of zip codes to grid cells was performed using the zip code centroids, and thus each rectangular region on the grid was assumed to correspond to the set of zip codes with centroid coordinates contained in that rectangle. A 1-day temporal window ($W = 1$) was used for all of these runs. We used a 28-day moving average (MA28) to compute expected counts, and the EBP statistic to detect clusters. For each of the three simulations discussed above (assuming small, medium, and large outbreaks respectively), we

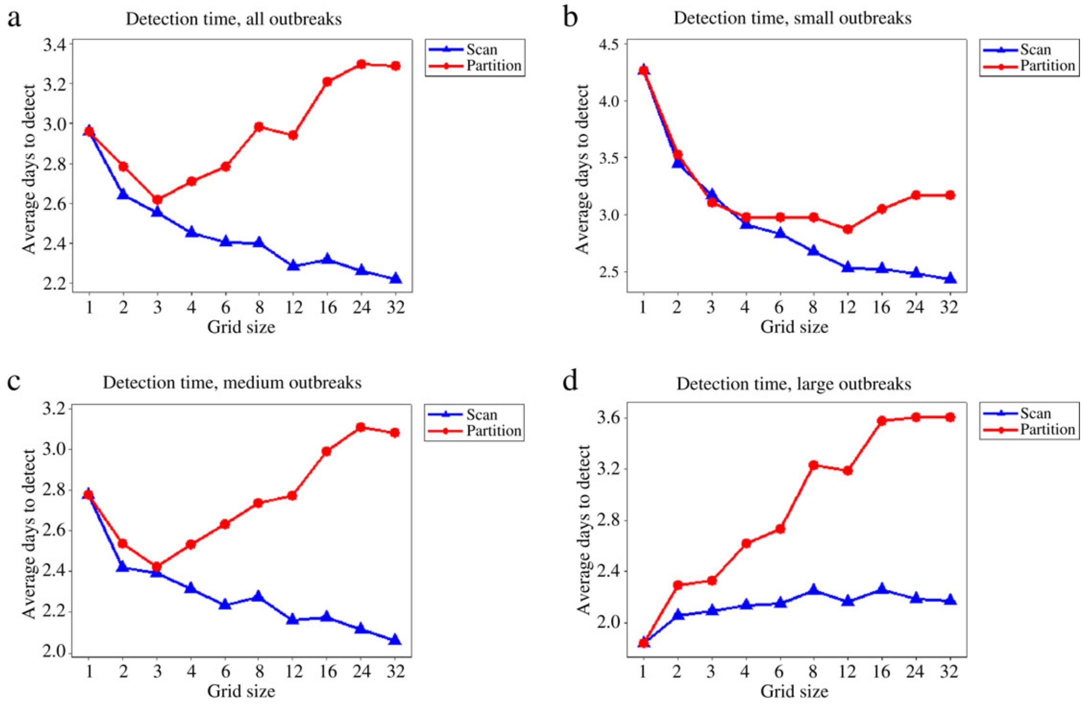

Fig. 1. A comparison of the scan statistic and fixed partition methods. Average days to detection at 1 false positive per month, as a function of grid size. (a) shows the average performance over all outbreaks, and (b)–(d) show the average performance for small, medium, and large outbreaks respectively.

computed the average detection time (with a 7-day penalty for undetected outbreaks, as discussed above) at a fixed false positive rate of 1/month.

Fig. 1(a) compares the average detection time for the expectation-based scan statistic and the fixed partition approach, averaged over all three outbreak simulations. We observe that the scan statistic outperforms fixed partitions for all grid sizes $N > 1$, with differences of over 1 day for large grid sizes. Both methods required 2.96 days to detect for $N = 1$. As expected, the performance of the scan statistic approach improved with an increased grid size, achieving the fastest detection (2.22 days to detect) at the largest grid size evaluated ($N = 32$). On the other hand, the performance of the fixed partition method improved only up to $N = 3$ (2.62 days to detect), and then declined rapidly for increasing grid sizes (up to 3.29 days to detect for $N = 32$).

To better understand these performance differences, we compare the average detection times separately for small, medium, and large outbreak sizes in Fig. 1(b)–(d) respectively. For small and medium outbreaks, the performance of the scan statistic approach improved with increasing grid sizes, while the performance of the fixed partition approach was optimized for intermediate grid sizes ($N = 12$ for small outbreaks and $N = 3$ for medium outbreaks). For large outbreaks, both the fixed partition and scan statistic methods performed best for $N = 1$ (1.84 days to detect). However, the performance of fixed partitions deteriorated rapidly with increasing grid size (up to 3.61 days to detect for $N = 32$), while the performance of the scan statistic remained approximately constant for $N \geq 2$ (requiring 2.06 days to detect for $N = 2$, and 2.17 days to detect for $N = 32$).

In Fig. 2, we compare the AMOC curves for the scan statistic and fixed partition approaches, using the optimal grid size for each method. We chose the grid size with the lowest detection time, at 1 false positive per month (i.e. $N = 32$ for the scan statistic, and $N = 3$ for fixed partitions), and then compared the detection times for all false positive rates between 1

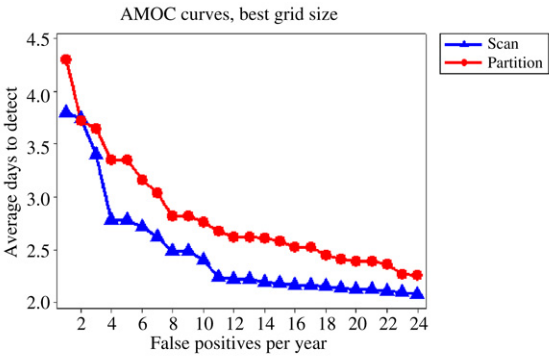

Fig. 2. AMOC curves for the scan statistic and fixed partition methods. Average days to detection, as a function of the false positive rate. A grid size of $N = 32$ is used for the scan statistic, and $N = 3$ for the fixed partition.

and 24 fp/year. These curves demonstrate consistently faster detection for the scan statistic than for fixed partitions, and all of these differences (with the single exception of 2 fp/year) were found to be statistically significant ($p < 10^{-6}$). These results demonstrate that the performance improvements shown in Fig. 1 are not simply an artifact of the chosen false positive rate.

These results demonstrate the primary advantage of the scan statistic over typical, fixed partition approaches to time series monitoring: by aggregating information across multiple spatial locations (and thus scanning over both large and small spatial regions), scan statistics can achieve high detection performance for all outbreak sizes. The only requirement is a sufficiently high spatial resolution: in our example, a grid size of $N = 12$ was sufficient to achieve a high average performance, and further (but slight) improvements were seen for higher resolutions. The fixed partition approach, on the other hand, must explicitly trade off the detection performance for small and large outbreaks, since coarse spatial resolutions (small N) and fine spatial resolutions (large N) have low detection performances for small and large outbreaks respectively.

A second advantage of the scan statistic approach as compared to fixed partitions is a more accurate identification of which spatial locations are affected by the outbreak, since the scan statistic can detect a collection of grid cells, while fixed partitions can only detect a single grid cell. To compare the spatial accuracy of the scan statistic and fixed partition methods, we computed the spatial precision and spatial recall of each method as a function of grid size, and then computed the F-measure (harmonic mean of precision and recall). The overall spatial accuracy (as given by the F-measure) is shown in Fig. 3(a), while precision and recall are shown separately in Fig. 3(b) and (c). For the scan statistic, the F-measure increases with increasing grid size, up to a maximum of 67.3% at $N = 32$. For the fixed partition approach, however, the F-measure only increases through $N = 4$, reaching a maximum of 45.5%, and then deteriorates with increasing N. For both approaches, the precision tends to increase and recall tends to decrease with increasing N. However, recall decreases dramatically for fixed partitions (19% recall for $N = 32$), while it levels off for the scan statistic (83% recall for $N = 32$). For a given N, the fixed partition approach has consistently higher precision but much lower recall than the scan statistic, since it identifies the single 1×1 grid cell which has been most affected by the outbreak, rather than the collection of all grid cells containing affected locations.

While the scan statistic demonstrates clear advantages over the fixed partition approach in terms of its ability to detect and localize outbreaks, it is much more computationally expensive, because it requires a search over all rectangular regions on the grid, rather than all grid cells. An exhaustive search over all rectangular regions requires $O(N^4)$ time for an $N \times N$ grid, while searching over grid cells only requires $O(N^2)$. However, for large N, many grid cells may be empty, and thus many rectangles may contain identical sets of spatial locations. For a fixed set of spatial locations (e.g. zip code centroids), we can compute the set of locations contained in each region in advance, and remove duplicate regions to reduce the computation time. To quantify the differences in run time between the scan statistic and fixed partition approaches, we ran each method (with a grid size varying from $N = 1$ to $N = 32$) on the 281 days of ED data from Allegheny County. Fig. 4 shows the total run time in seconds for each method and grid size. The run time for the fixed partition methods was dominated by the cost of loading the ED data, and thus remained approximately constant (3.5–3.6 s) as the grid size increased. The run time for the scan statistic increased by a factor of 4-7x for each doubling of grid size, reaching a maximum of 423.2 s for $N = 32$.

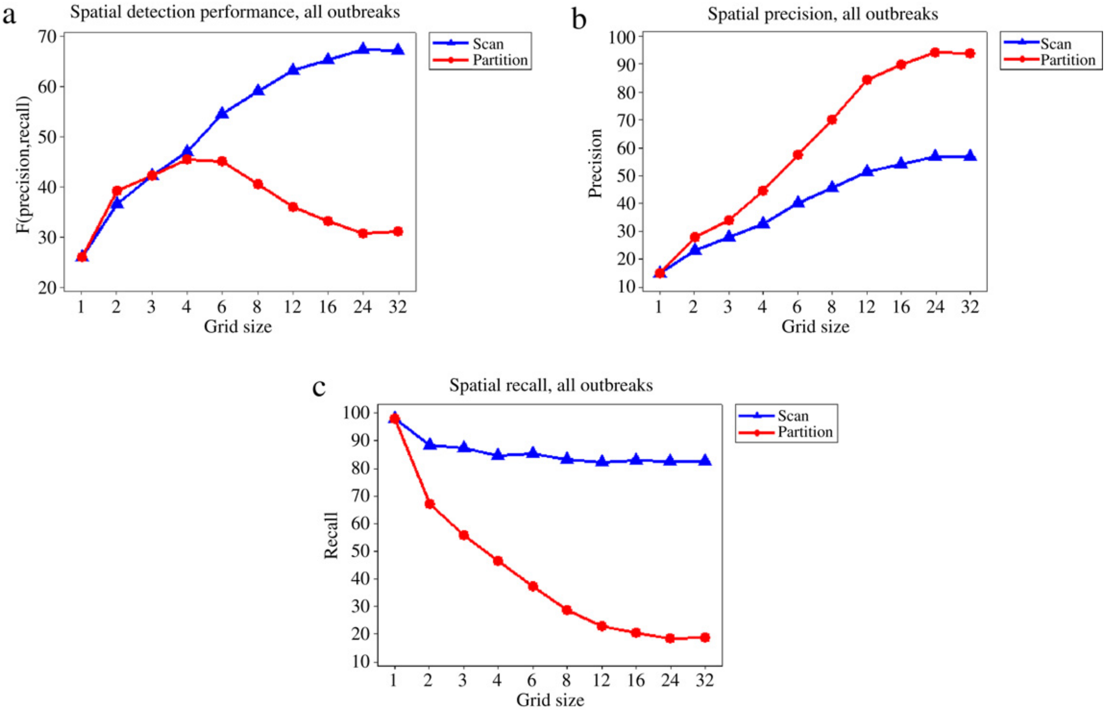

Fig. 3. A comparison of the scan statistic and fixed partition methods. Spatial detection accuracy at day 4 of the outbreak, as a function of grid size. (a) shows the F-measure (harmonic mean of spatial precision and recall). (b) and (c) show precision and recall respectively.

Thus, computational efficiency can become a factor in choosing which grid size to use, especially when the number of spatial locations is large or many days of data must be examined. For county-level data, a grid size of $N = 12$ or $N = 16$ appears to be a good tradeoff between detection power and computation time. For larger datasets (e.g. nationwide sales of over-the-counter medications), we must use much larger grid sizes, and thus more efficient algorithms are required to make the scan statistic computationally feasible. As noted above, Neill and Moore (2004) and Neill et al. (2005) present a "fast spatial scan" algorithm which can be used to speed up the scan statistic by 2–3 orders of magnitude for large grid sizes, with no loss of accuracy.

5.2. *Comparison of window sizes*

In our second set of experiments, we examined how the performance of the expectation-based scan statistic varies with the maximum temporal window size W, where we search over space–time regions with durations between 1 and W. We evaluated a range of temporal window sizes $W = 1$ to $W = 7$, using a fixed grid size of $N = 16$. As in the previous experiments, we used the 28-day moving average (MA28) to compute expected counts, and the EBP statistic to detect clusters. For each of the three simulations discussed above (assuming small, medium, and large outbreaks respectively), we computed the average detection time (with a penalty for undetected outbreaks, as discussed above) for false positive rates varying from 1–24 fp/year. We averaged detection times over all 3000 simulated outbreaks to form an AMOC curve (measuring the average detection time as a function of the false positive rate), as shown in Fig. 5(a). Because the relative performances of different temporal window sizes is strongly dependent on the temporal progression of the outbreak, we also ran an additional simulation (1000 randomly generated outbreaks) of "gradual" outbreaks, with a slower increase in the injected counts ($\Delta = 1$) and a 28-day duration. These outbreaks affected 10–20 zip codes surrounding a randomly

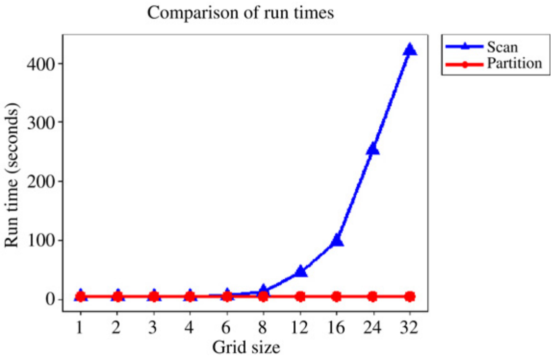

Fig. 4. A comparison of the scan statistic and fixed partition methods. Run time in seconds, as a function of grid size.

selected center zip code, as in the "medium" outbreak simulation. AMOC curves for these gradual outbreaks are shown in Fig. 5(b).

In Fig. 5(a), we observe that the relative performance of different temporal window sizes (as measured by the average detection time) is highly dependent on the allowable false positive rate. For very low false positive rates (1–3 false positives per year), longer temporal windows ($W > 1$) detect between 0.5 and 0.8 days faster than a 1-day temporal window. However, for higher false positive rates, the fastest detection time is achieved for $W = 1$, and longer window sizes detect between 0.15 and 0.25 days slower. For example, at a fixed false positive rate of 1 fp/month, $W = 1$ detects outbreaks in an average of 2.32 days, while methods with longer W required between 2.49 and 2.56 days. Detection time tended to increase with window size for $W > 3$: since the number of injected cases increased sharply, most outbreaks were detected by the third outbreak day, and thus longer window sizes did not improve performance. For more gradual outbreaks (Fig. 5(b)), we again observe a strong dependence of performance on the allowable false positive rate. In this case, we saw huge performance benefits for longer temporal windows at low false positive rates. At 1 fp/year, $W = 1$ required an average of 23.4 days to detect. Detection time improved to 12.3 days for $W = 2$, 10.2 days for $W = 3$, and continued to improve with increasing temporal window size (8.5 days to detect at $W = 7$). Longer windows improved performance for false positive rates up to 8 fp/year, but the 1-day window performed best for false positive rates above 10 fp/year, typically outperforming longer windows by 0.1–0.4 days.

Thus, longer temporal window sizes improve the performance when the allowable false positive rate is low, and when the outbreak is more gradual; on the other hand, a 1-day window is best for quickly growing outbreaks and high allowable false positive rates. The signal to noise plots shown in Fig. 6(a) and (b) (for standard and gradual outbreaks respectively) help to explain these performance results. The solid line in each graph represents the mean score on each day of the outbreak (averaged over all simulated outbreaks), while the dashed lines represent the 95th and 99th percentiles of the background score distribution (i.e., the scores computed for all 281 days of baseline data with no outbreaks injected). We observe that the score increases more rapidly over the course of the outbreak for longer window sizes, but

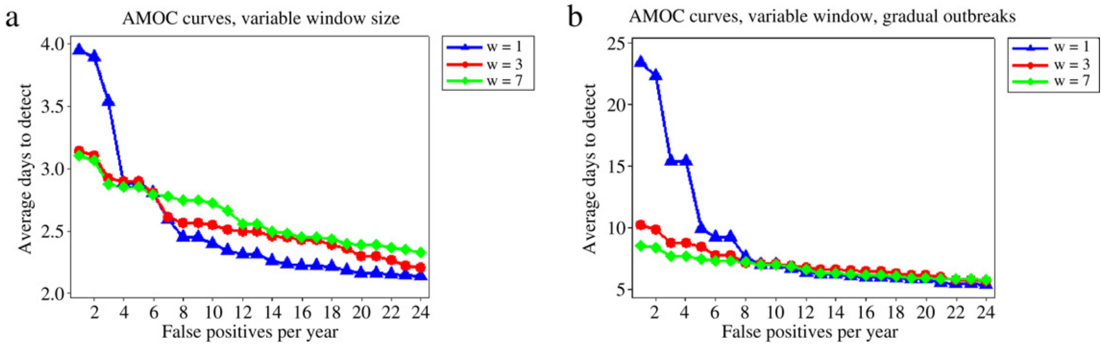

Fig. 5. AMOC curves for scan statistic methods with varying temporal window size ($W = 1$, $W = 3$, and $W = 7$). Average days to detection, as a function of the false positive rate. (a) compares the performance based on the original, 7-day outbreak simulations. (b) compares the performance using a more gradual, 28-day outbreak simulation.

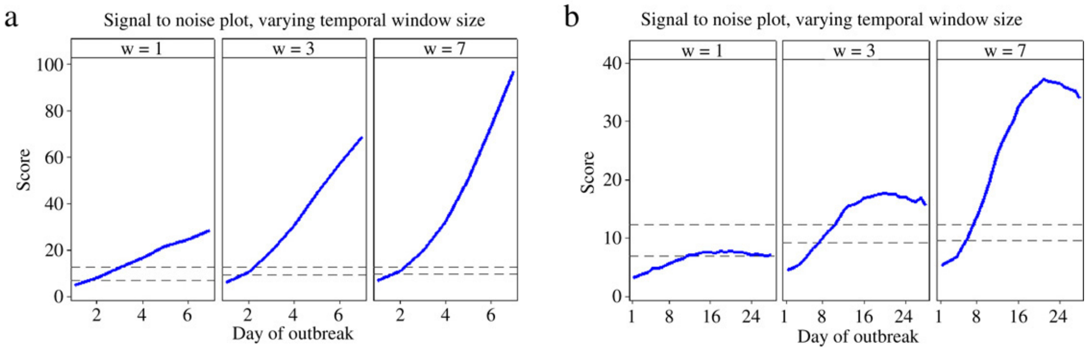

Fig. 6. Signal to noise plots for scan statistic methods with varying temporal window sizes ($W = 1$, $W = 3$, and $W = 7$). The thick line in each plot is the average score on each day of the outbreak. The dashed lines are the 95th and 99th percentiles of the score distribution for the baseline data with no outbreaks injected. (a) compares the performance on the "medium" 7-day outbreak simulation. (b) compares the performance on the "gradual" 28-day outbreak simulation.

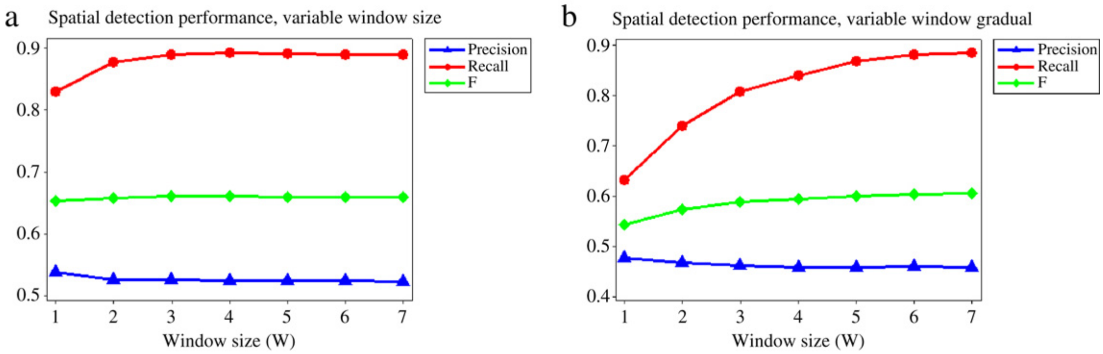

Fig. 7. Spatial precision, recall, and F-measures for scan statistic methods at the midpoint of an outbreak, as a function of temporal window size ($W = 1 \ldots 7$). (a) compares the performance based on the original, 7-day outbreak simulations. (b) compares the performance using a more gradual, 28-day outbreak simulation.

scores are also higher for the background data (as evident from the larger 95th percentile value; the 99th percentile values remained approximately constant). By the third day of the rapidly growing outbreaks (Fig. 6(a)), the longer windows had scores well above the 99th percentile value, while $W = 1$ had a score comparable to the 99th percentile value, explaining its lower performance for low false positive rates. On the other hand, $W = 1$ was faster to reach the 95th percentile value, explaining its improved performance for high false positive rates. For the gradual outbreaks (Fig. 6(b)), the longer windows had scores well above the 99th percentile value by the tenth outbreak day, while the scores for $W = 1$ never approach this value. This explains the extremely poor performance of $W = 1$ for the gradually growing outbreaks, and the large performance gains for longer window sizes.

We also compare the spatial detection accuracy for varying window sizes, measuring the average spatial precision and spatial recall for each method at the midpoint of the outbreak, and computing the F-measure (harmonic mean of precision and recall). For the rapidly growing outbreaks (Fig. 7(a)), we observe that the longer window sizes have somewhat higher recall and slightly lower precision, resulting in an slightly increased F-measure. Recall increased from 83.0% for $W = 1$ up to 89.2% for longer windows. For the gradually growing outbreaks (Fig. 7(b)), recall improved substantially with increasing window sizes, from 63.0% for $W = 1$ up to 88.5% for $W = 7$. Precision stayed approximately constant, and thus the F-measure also increased with increasing window size.

These results demonstrate two advantages of using a longer temporal window size: improved detection time for lower false positive rates, and improved spatial detection accuracy. Both of these advantages are largest when the outbreak emerges gradually over time; for more rapidly emerging outbreaks, a 1-day temporal window is sufficient and may even achieve slightly faster detection.

5.3. Comparison of time series analysis methods

In our third set of experiments, we compared the 28-day moving average (MA28) to seven other time series analysis methods: 7-day, 14-day, and 56-day moving averages (MA7, MA14, MA56), moving averages with global and local day-of-week adjustments (MAGD, MALD), the multiplicative Holt–Winters' method (HW), and the "current day" method (CD). All of these time series analysis methods are described in Section 3. For each of the three simulations discussed above (assuming small, medium, and large outbreaks respectively), we computed the average detection time (with a penalty for undetected outbreaks, as discussed above) for false positive rates varying from 1–24 fp/year. A grid size of $N = 16$, and a 1-day temporal window ($W = 1$) were used for all of these runs, and EBP was used to detect clusters. We averaged results for the three simulations, producing the AMOC curves shown in Fig. 8.

From Fig. 8(a), we observe that MA28 tends to outperform the MA7 and MA14 methods (particularly at low false positive rates), while MA28 and MA56 achieve similar performances. At a fixed false positive rate of 1 fp/month, MA14 and MA56 had similar detection times to MA28 (2.37, 2.34, and 2.32 days to detect respectively), while MA7 required significantly longer (2.65 days to detect). MA56 also had a slightly improved F-measure as compared to MA28 (66.6% vs. 65.4%), while MA7 and MA14 had significantly lower F-measures (55.2% and 61.8% respectively).

These results suggest that at least 28 days of data are needed to accurately estimate baselines for the expectation-based scan statistic. Using fewer days to compute baselines leads to higher variance of the estimates. Additionally, baseline estimates during the outbreak are biased upward because they incorporate the injected counts into the estimate, and this bias is larger when fewer days of data are used. (Alternatively, the most recent 7 or 14 days of data can be omitted when including baselines, but we do not examine this approach here.) Using fewer days of data can be advantageous in application domains with strong temporal trends, since in these cases the most recent data is a more accurate predictor than the past data. However, the poor performance of MA7 suggests that these trends are not present in the ED dataset.

Fig. 8(b) examines whether adjustment for day-of-week patterns improves detection performance by comparing 28-day moving averages without day-of-week adjustment (MA28), with global day-of-week adjustment (MAGD), and with local day-of-week adjustment (MALD). We observe that, while MAGD and MALD achieve faster detection than MA28 for very low false positive rates (1–3 fp/year), MA28 consistently outperforms the other two methods for higher false positive rates. At a fixed false positive rate of 1/month, MA28 detected outbreaks in an average of 2.32 days, as compared to 2.52 days for MAGD and 2.60 days for MALD. Additionally, while MA28 and MAGD had very similar spatial detection accuracy (F-measure 65.4% for both methods), MALD had lower spatial accuracy (F-measure 61.6%).

These results suggest that, for the ED dataset, no substantial performance benefits are gained by adjusting for day-of-week effects. The improved performance of the day-of-week adjustments at low false positive rates suggests that some weekly trends are present in the data, and the improved performance of MALD over MAGD (for 1–3 fp/year) also suggests that some space by day-of-week interaction is present. However, for higher false positive rates these benefits are outweighed by the increased variance of the baseline estimates resulting from day-of-week adjustment. MALD in particular suffered decreased precision and recall as compared to MA28, suggesting that our attempts to infer each location's weekly trends resulted in high variance due to the small number of counts.

Finally, we compared the MA28 method to two previously proposed methods for time series analysis: the multiplicative Holt–Winters' (HW) method, an exponential smoothing method that accounts for both (linear) seasonal and (cyclical) day-of-week trends; and the "current day" (CD) method used

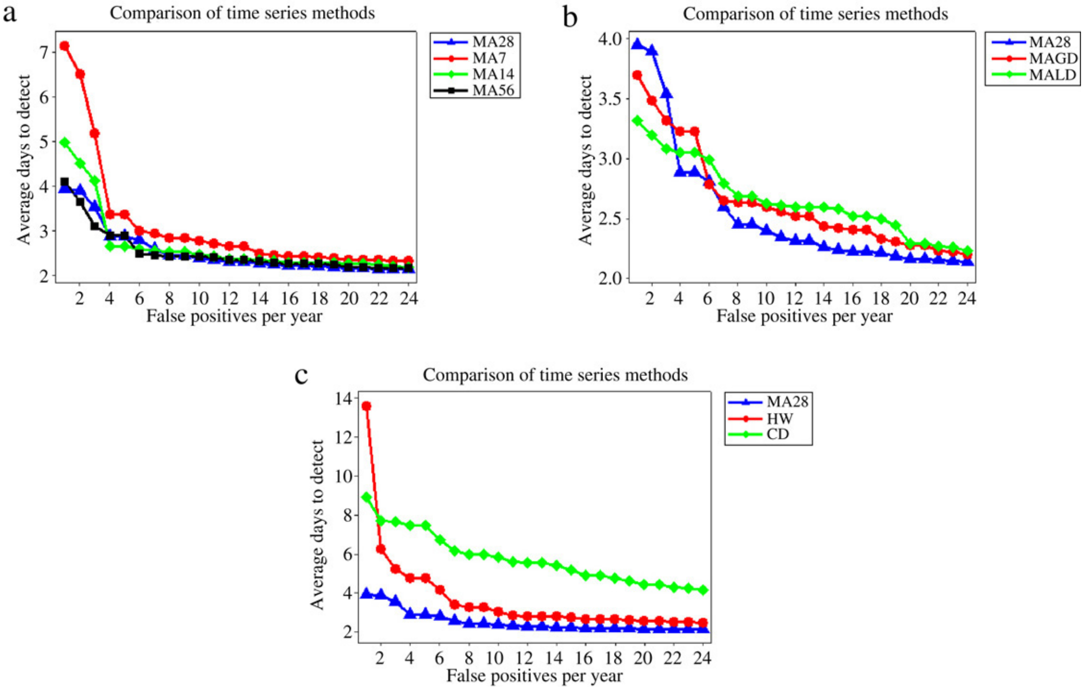

Fig. 8. AMOC curves for the expectation-based scan statistic, using different time series analysis methods to obtain expected counts. Average days to detection, as a function of the false positive rate. (a) compares the 28-day moving average method (MA28) to 7-day, 14-day, and 56-day moving averages. (b) compares MA28 to 28-day moving average methods globally and locally adjusted for day of the week (MAGD and MALD). (c) compares MA28 to the Holt–Winters' method (HW) and the current day method (CD).

in the space–time permutation statistic (Kulldorff et al., 2005), which adjusts for the global aggregate count and thus detects only space–time interaction, rather than purely spatial or purely temporal trends. From Fig. 8(c), we observe that MA28 consistently outperforms both methods, detecting approximately four days faster than CD across the entire range of false positive rates. The HW method detects approximately 0.5 days slower than MA28 for high false positive rates, but performs extremely poorly for low false positive rates (ten days slower than MA28, and five days slower than CD, at 1 fp/year). At a fixed false positive rate of 1/month, HW and CD underperformed MA28, detecting outbreaks in 2.82 and 5.55 days respectively (as compared to 2.32 for MA28). The performance of CD was particularly poor for large outbreaks (requiring 10.58 days to detect at 1 fp/month), and was at least 0.4 days slower than MA28 for all outbreak sizes. Additionally, the spatial accuracy for HW and CD was low (F-measures of 56.7% and 57.8% respectively, as compared to 65.4% for MA28).

These results suggest that adjustments for day-of-week and seasonality are not necessary in the ED data, and that the use of more complicated methods such as Holt–Winters' for this dataset may lead to unstable baseline estimates, and thus increase the number of false positives. For our data, a simple 28-day moving average was shown to be sufficiently accurate, achieving rapid and accurate detection of outbreaks. Other datasets may require the use of more complex time series analysis methods to account for seasonal and day-of-week trends: for example, over-the-counter medication sales often exhibit substantial weekly and seasonal variation.

5.4. *Comparison of expectation-based and Kulldorff scan statistics*

In our fourth set of experiments, we compared our expectation-based Poisson (EBP) scan statistic

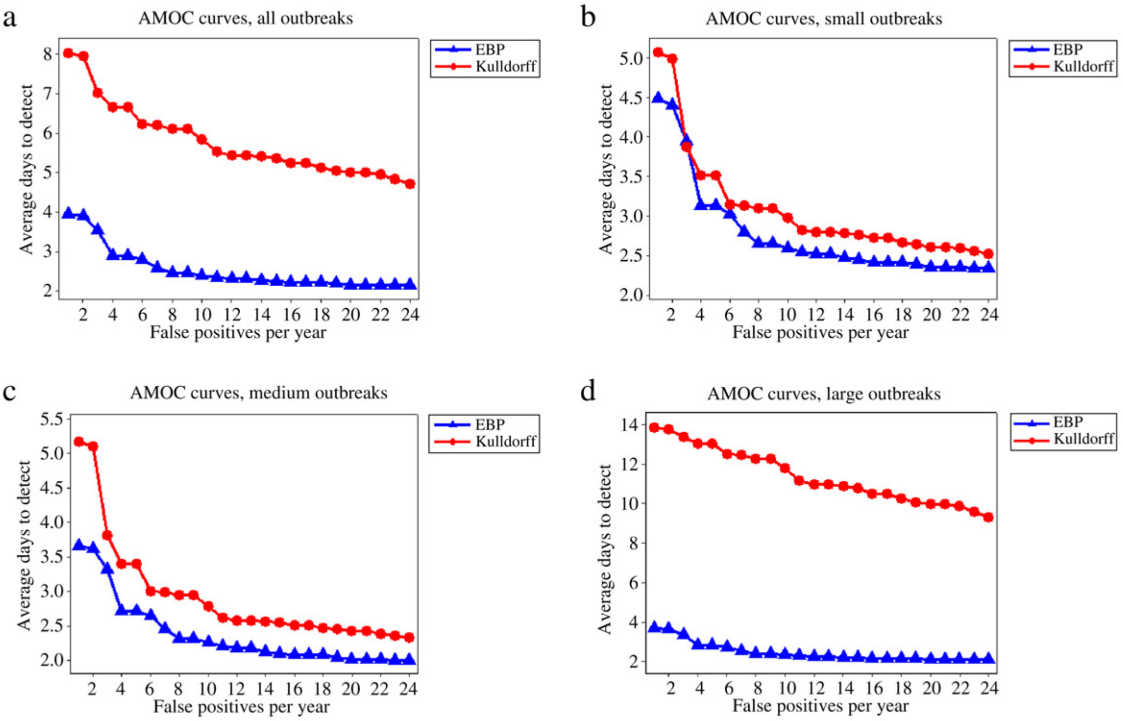

Fig. 9. AMOC curves for the expectation-based Poisson (EBP) and Kulldorff scan statistic methods. Average days to detection, as a function of the false positive rate. (a) shows the average performance over all outbreaks, and (b)–(d) show the average performance for small, medium, and large outbreaks respectively.

method discussed above to Kulldorff's original space–time scan statistic (Kulldorff, 2001). Both of these methods are described in Section 4. For each of the three simulations discussed above (assuming small, medium, and large outbreaks respectively), we computed the average detection time (with a penalty for undetected outbreaks, as discussed above), for false positive rates varying from 1–24 fp/year, producing the AMOC curves shown in Fig. 9. Fig. 9(a) shows the mean detection times averaged over all three outbreak simulations, while Fig. 9(b)–(d) show the mean detection times for small, medium, and large outbreaks respectively. A grid size of $N = 16$, and a 1-day temporal window ($W = 1$) were used for all of these runs; baselines were estimated using a 28-day moving average (MA28).

From Fig. 9(a), we observe that our expectation-based scan statistic outperforms Kulldorff's statistic by a large margin across all false positive rates, achieving over three days faster detection. At a fixed false positive rate of 1 fp/month, EBP detected outbreaks in an average of 2.32 days, as compared to 5.46 days for Kulldorff's statistic. Fig. 9(b)–(d) reveal that the performance differences are much smaller for small and medium outbreaks, though EBP outperformed Kulldorff's statistic across all three outbreak sizes. At a fixed false positive rate of 1/month, EBP detected 0.28 days faster than Kulldorff's statistic for small outbreaks (2.52 vs. 2.80 days to detect), 0.41 days faster for medium outbreaks (2.17 vs. 2.58 days to detect), and 8.73 days faster for large outbreaks (2.26 vs. 10.99 days to detect). All of these results for 1 fp/month were found to be statistically significant ($p < 0.001$). The performance difference for large outbreaks is particularly striking: while EBP detected 100% of these outbreaks, Kulldorff's statistic had a detection rate of only 30.2%, resulting in an average detection time that was longer than the outbreak duration (because of the 7-day penalty for missed outbreaks).

We also computed the spatial detection accuracy for each method at the midpoint of the outbreak (day 4). The expectation-based scan statistic had much higher recall than Kulldorff's statistic (83.1% vs. 66.0%), though its precision was somewhat lower (53.9% vs. 58.4%). This resulted in an F-measure of 65.4%, as compared to 62.0% for Kulldorff's statistic, demonstrating improved spatial accuracy. For large outbreaks, EBP achieved an average recall of 82.5% as compared to 31.4% for Kulldorff's statistic, demonstrating that it detects a much larger region containing most of the affected locations rather than only a small subset of the most affected locations.

These results confirm our expectations: since Kulldorff's statistic compares the relative risks inside and outside the search region rather than comparing the actual and expected counts inside the search region, it has low power to detect large outbreaks that affect many spatial locations. Somewhat surprisingly, the expectation-based Poisson statistic achieves significantly improved detection time for small to medium-sized outbreaks as well. As noted above, we expect that Kulldorff's statistic may outperform EBP for small outbreak sizes when we have poor estimates of the expected counts (e.g. failure to account for seasonal or day-of-week trends in datasets where these trends are present). However, these performance gains were not observed for the Emergency Department data, suggesting that our time series analysis is producing accurate baseline estimates.

6. Discussion

We have presented the expectation-based scan statistic, a method for monitoring multiple spatially localized time series in order to detect spatial clusters of increased counts. This method consists of two steps: estimating the expected counts for each spatial location for each recent day, and detecting space–time regions where the observed counts are significantly higher than expected. In Section 5.1, we demonstrated that the expectation-based scan statistic achieves significantly higher detection power than typical time series monitoring approaches (e.g. choosing a desired level of aggregation, and then monitoring each of the aggregated time series separately). By searching many overlapping sets of spatial locations with varying shapes and sizes, we can achieve high power to detect spatial patterns, whether they affect a single location, all locations, or a subset of locations. This is very different from the typical "fixed partition" approach, which loses power to detect small patterns if the space is coarsely partitioned, and loses power to detect large patterns if the space is finely partitioned. An additional advantage of the scan statistic over fixed partitions is higher spatial accuracy (improved ability to determine which spatial locations are affected by an event). This advantage is achieved by mapping the locations to a grid with fine spatial resolution and then detecting clusters of affected grid cells.

When using the expectation-based scan statistic framework for time series monitoring, we must answer four main questions: which set of spatial regions to search, which temporal window size W to choose, which time series analysis method to use for calculating expected counts, and which statistic to use for detecting clusters of higher than expected counts. We consider the set of search regions in Section 5.1. Our study assumes that locations are mapped to a uniform $N \times N$ grid, and we search over all rectangular regions on the grid. We demonstrate that a grid resolution of at least $N = 12$ is necessary for high detection power on county-level Emergency Department data; even higher grid resolutions can achieve further (but slight) improvements in detection time, but also substantially increase the computations required. We address the second question in Section 5.2, showing that the optimal setting of the temporal window size depends on both the nature of the outbreak (a rapid or gradual increase in counts) and the allowable false positive rate. We address the third question in Section 5.3, demonstrating that a simple 28-day moving average method is sufficient for rapid and accurate outbreak detection using our Emergency Department data. In other datasets with strong day-of-week or seasonal trends, other methods which account for these trends may achieve higher performance. Finally, we address the fourth question in Section 5.4, demonstrating that our expectation-based Poisson scan statistic outperforms the traditional Kulldorff space–time scan statistic approach.

We note that this work did not examine the impact of the choice of region shape on detection power, as this question has been addressed by a number of recent studies. While Kulldorff's original spatial scan statistic (Kulldorff, 1997) performed a search over the set of circular regions, many other shapes including rectangles (Neill et al., 2005b), ellipses (Kulldorff, Huang, Pickle, & Duczmal, 2006), and various sets of irregular regions (Duczmal & Assuncao, 2004; Patil & Taillie, 2004; Tango & Takahashi, 2005) have been considered. In general, the optimal region shape is strongly dependent on the outbreak's region of effect. Circular (or square) search regions perform well for compact clusters but poorly for elongated clusters, rectangular or elliptical search regions perform well for elongated clusters, and irregular search regions perform best when the cluster shape is highly irregular (Duczmal, Kulldorff, & Huang, 2006).

Additionally, many other variants of the scan statistic have recently been proposed, including Gaussian (Neill, 2006), robust (Neill & Sabhnani, 2007), model-adjusted (Kleinman, Abrams, Kulldorff, & Platt, 2005), nonparametric (Neill & Lingwall, 2007), and Bayesian (Neill, Moore, & Cooper, 2006, 2007) methods. We believe that some of these more complex methods may further improve detection performance, and we are in the process of conducting a large-scale evaluation of these methods using hospital Emergency Department and over-the-counter medication sales data. The preliminary results of this evaluation (Neill, 2007) complement the present work by demonstrating that the relative performance of different statistics is highly affected by the dataset characteristics (e.g. large or small daily counts, presence or absence of seasonal and day-of-week trends), as well as the characteristics of the injected outbreak.

References

Ailamaki, A., Faloutsos, C., Fischbeck, P., Small, M., & VanBriesen, J. (2003). An environmental sensor network to determine drinking water quality and security. *SIGMOD Record*, *32*(4), 47–52.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw-Hill.

Buckeridge, D. L., Burkom, H. S., Moore, A. W., Pavlin, J. A., Cutchis, P. N., & Hogan, W. R. (2004). Evaluation of syndromic surveillance systems: Development of an epidemic simulation model. *Morbidity and Mortality Weekly Report*, *53*, 137–143 (Supplement on Syndromic Surveillance).

Burkom, H. S., Murphy, S. P., Coberly, J., & Hurt-Mullen, K. (2005). Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report*, *54*, 55–62. (Supplement on Syndromic Surveillance).

Burkom, H. S., Murphy, S. P., & Shmueli, G. (2007). Automated time series forecasting for biosurveillance. *Statistics in Medicine*, *26*(22), 4202–4218.

Corcoran, J. J., Wilson, I. D., & Ware, J. A. (2003). Predicting the geo-temporal variations of crime and disorder. *International Journal of Forecasting*, *19*, 623–634.

Davies, R., for the ECADS partners and collaborator (2006). Detection of Walkerton gastroenteritis outbreak using syndromic surveillance of emergency room records. *Advances in Disease Surveillance*, *1*, 20.

Duczmal, L., & Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, *45*, 269–286.

Duczmal, L., Kulldorff, M., & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal Computational & Graphical Statistics*, *15*(2), 428–442.

Fawcett, T., & Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the 5th international conference on knowledge discovery and data mining* (pp. 53–62).

Gorr, W., & Harries, R. (2003). Introduction to crime forecasting. *International Journal of Forecasting*, *19*, 551–555.

Hjalmars, U., Kulldorff, M., Gustafsson, G., & Nagarwalla, N. (1996). Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, *15*, 707–715.

Hotelling, H. H. (1947). Multivariate quality control. In C. Eisenhart, M. W. Hastay, & W. A. Wallis (Eds.), *Techniques of statistical analysis* (pp. 111–184). New York: McGraw-Hill.

Kleinman, K., Abrams, A., Kulldorff, M., & Platt, R. (2005). A model-adjusted space–time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, *133*(3), 409–419.

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

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, *164*, 61–72.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., & Key, C. (1998). Evaluating cluster alarms: A space–time scan statistic and cluster alarms in Los Alamos. *American Journal of Public Health*, *88*, 1377–1380.

Kulldorff, M., Feuer, E. J., Miller, B. A., & Freedman, L. S. (1997). Breast cancer clusters in the northeast United States: A geographic analysis. *American Journal of Epidemiology*, *146*(2), 161–170.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., & Mostashari, F. (2005). A space–time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine*, *2*(3), e59.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, *25*, 3929–3943.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, *14*, 799–810.

Levine, N. (1999). *CrimeStat: A spatial statistics program for the analysis of crime incident locations*. National Institute of Justice, Washington, DC.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., & Kulasekera, V. (2003). Dead bird clustering: A potential early warning system for West Nile virus activity. *Emerging Infectious Diseases*, *9*, 641–646.

Neill, D. B. (2006). *Detection of spatial and spatio-temporal clusters*. Tech. rep CMU-CS-06-142. Ph.D. thesis. Carnegie Mellon University, Department of Computer Science.

Neill, D. B. (2007). An empirical comparison of spatial scan statistics for outbreak detection. *Advances in Disease Surveillance*, *4*, 259.

Neill, D. B., & Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, *4*, 106.

Neill, D. B., & Moore, A. W. (2005). Anomalous spatial cluster detection. In *Proceedings of KDD 2005 workshop on data mining methods for anomaly detection* (pp. 41–44).

Neill, D. B., Moore, A. W., & Cooper, G. F. (2006). A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems*, *18*, 1003–1010.

Neill, D. B., Moore, A. W., & Cooper, G. F. (2007). A multivariate Bayesian scan statistic. *Advances in Disease Surveillance*, *2*, 60.

Neill, D. B., Moore, A. W., & Sabhnani, M. R. (2005a). Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report*, *54*, 197. (Supplement on Syndromic Surveillance).

Neill, D. B., Moore, A. W., Sabhnani, M. R., & Daniel, K. (2005b). Detection of emerging space–time clusters. In *Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 218–227).

Neill, D. B., & Sabhnani, M. R. (2007). A robust expectation-based spatial scan statistic. *Advances in Disease Surveillance*, *2*, 61.

Neill, D. B., & Moore, A. W. (2004). Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 256–265).

Neill, D. B., Moore, A. W., Pereira, F., & Mitchell, T. (2005). Detecting significant multidimensional spatial clusters. *Advances in Neural Information Processing Systems*, *17*, 969–976.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*, 100–115.

Patil, G. P., & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, *11*, 183–197.

Sabhnani, M. R., Neill, D. B., Moore, A. W., Tsui, F. C., Wagner, M. M., & Espino, J. U. (2005). Detecting anomalous clusters in pharmacy retail data. In *Proceedings of the KDD 2005 workshop on data mining methods for anomaly detection* (pp. 58–61).

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York: Van Nostrand.

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, *4*, 11.

Trigg, D. W. (1964). Monitoring a forecasting system. *Operations Research Quarterly*, *15*, 271–274.

Wallstrom, G. L., Wagner, M. M., & Hogan, W. R. (2005). High-fidelity injection detectability experiments: A tool for evaluation of syndromic surveillance systems. *Morbidity and Mortality Weekly Report*, *54*, 85–91. (Supplement on Syndromic Surveillance).