Research Article

Statistics in Medicine

Received 22 January 2010, Accepted 22 January 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3881

Fast Bayesian scan statistics for multivariate event detection and visualization

Daniel B. Neill[*†]

The multivariate Bayesian scan statistic (MBSS) is a recently proposed, general framework for event detection and characterization in multivariate space–time data. MBSS integrates prior information and observations from multiple data streams in a Bayesian framework, computing the posterior probability of each type of event in each space–time region. MBSS has been shown to have many advantages over previous event detection approaches, including improved timeliness and accuracy of detection, easy interpretation and visualization of results, and the ability to model and accurately differentiate between multiple event types. This work extends the MBSS framework to enable detection and visualization of irregularly shaped clusters in multivariate data, by defining a hierarchical prior over all subsets of locations. While a naive search over the exponentially many subsets would be computationally infeasible, we demonstrate that the total posterior probability that each location has been affected can be efficiently computed, enabling rapid detection and visualization of irregular clusters. We compare the run time and detection power of this 'Fast Subset Sums' method to our original MBSS approach (assuming a uniform prior over circular regions) on semi-synthetic outbreaks injected into real-world Emergency Department data from Allegheny County, Pennsylvania. We demonstrate substantial improvements in spatial accuracy and timeliness of detection, while maintaining the scalability and fast run time of the original MBSS method.

Keywords: event detection; disease surveillance; scan statistics

1. Introduction

This work focuses on the task of *disease surveillance*, in which we monitor electronically available public health data sources, such as hospital visits and medication sales, to automatically detect emerging outbreaks of disease. Our goal is to develop disease surveillance systems which can identify outbreaks in the very early stages (typically, before a definitive diagnosis of the outbreak disease can be obtained), enabling a timely and effective public health response.

The multivariate Bayesian scan statistic (MBSS) [1] is a recently proposed Bayesian framework for detection and characterization of emerging outbreaks. MBSS integrates information from multiple data streams, enabling more timely and more accurate event detection, and can model and differentiate between multiple event types. MBSS can be used to detect emerging outbreaks of disease, pinpoint the affected spatial region, distinguish between different outbreak diseases, and reduce the number of 'false-positive' alerts due to irrelevant anomalies in the data [1]. By detecting and characterizing outbreaks in their early stages, MBSS can provide public health users with sufficient situational awareness to enable a rapid and informed response.

MBSS has been shown to have numerous advantages over the frequentist spatial scan statistics approach [1]. By integrating information from multiple data streams and incorporating prior knowledge of an outbreak's effects on the monitored data streams, MBSS can achieve more timely and more accurate detection, detecting an average of 1.3 days faster than Kulldorff's multivariate scan statistic [2]. MBSS can model and accurately differentiate between multiple event types, thus distinguishing between patterns due to different outbreak diseases and those due to other irrelevant events in the data. Outbreak models can be specified by a domain expert or learned automatically from a small number of labeled training examples. Randomization testing is not necessary in the Bayesian framework: since 999 or more Monte Carlo replications must typically be performed to obtain accurate p-values for the frequentist approach, we can obtain a 1000× speedup by avoiding the need for randomization. Finally, the results produced by MBSS (the posterior probability

H.J. Heinz III College, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.
[*]Correspondence to: Daniel B. Neill, H.J. Heinz III College, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.
[†]E-mail: neill@cs.cmu.edu

that each event type has affected each spatial region) are easy to interpret, visualize, and use for decision-making. MBSS computes the posterior probability $Pr(H_1(S, E_k)|D)$ that each outbreak type E_k has affected each spatial region S, and these results can be visualized as a 'posterior probability map', showing the total probability

$$Pr(H_1(s_i)|D) = \sum_{E_k} \sum_{S:s_i \in S} Pr(H_1(S, E_k)|D)$$

that each location s_i has been affected.

The primary limitation of MBSS is a computational issue common to many spatial scan statistic methods: to compute the posterior probabilities, we must evaluate a huge number of spatial regions S. If we place no constraints on the search region, all 2^N subsets of the N locations must be considered, which is computationally infeasible for even moderately large N. MBSS, like other typical scan statistic approaches, solves this problem by only considering a reduced set of regions, placing restrictions on the region shape. As in Kulldorff's original spatial scan statistic approach [3], MBSS restricts the search space to the set of $O(N^2)$ distinct circular regions centered at one of the N spatial locations. While this limitation of the search space makes MBSS computationally feasible, the assumption of circular outbreaks causes it to lose detection power for elongated or irregular outbreaks, as well as reducing its ability to accurately pinpoint the affected region.

Here we propose a new extension of the MBSS framework, 'Fast Subset Sums' (FSS), which enables timely and accurate detection and visualization of irregularly shaped clusters in multivariate data. FSS can efficiently compute the posterior probability map, as well as the total posterior probability of an outbreak, without computing the individual posterior probabilities of each spatial region. It defines a hierarchical prior distribution over regions which assigns non-zero prior probability to each of the 2^N subsets of locations, and efficiently computes the summed posterior probabilities using this distribution.

2. Methods

We briefly review the recently proposed MBSS framework, discuss its limitations, and then present our new 'FSS' method.

2.1. Multivariate Bayesian scan statistics

In the multivariate disease surveillance problem, our goal is to detect and characterize outbreaks based on their effects on the monitored data sources. We typically monitor aggregated count data for multiple spatial locations, time steps, and data streams. For example, to detect an influenza outbreak, we could monitor hospital Emergency Department (ED) visits, with each data stream representing the number of ED visits with a different symptom type (e.g. 'cough' or 'fever').

The MBSS framework [1] assumes a dataset D consisting of multiple data streams D_m, for $m = 1 \ldots M$. Each data stream consists of spatial time series data collected at a set of spatial locations s_i, for $i = 1 \ldots N$. For each stream D_m and location s_i, we have a time series of observed counts $c^t_{i,m}$ and a time series of expected counts $b^t_{i,m}$, where $t = 0$ represents the current time step and $t = 1 \ldots T$ represent the counts from 1 to T time steps ago respectively. For example, a given observed count $c^t_{i,m}$ might represent the total number of ED visits with fever symptoms, for a given zip code on a given day. The corresponding expected count $b^t_{i,m}$ would be the expected number of ED visits with fever symptoms in that zip code on that day, estimated from the time series of historical fever counts for that zip code. Here we calculate expected counts using a simple, 28-day moving average, but other time series analysis methods can be also be applied, in order to adjust for seasonal and day-of-week trends.

MBSS is designed to detect emerging outbreaks, identify the type of outbreak (e.g. distinguishing between anthrax and influenza), and pinpoint the affected locations. To do so, it compares the set of alternative hypotheses $H_1(S, E_k)$, each representing the occurrence of some outbreak type E_k in some subset of locations S, against the null hypothesis H_0 that no outbreaks have occurred. These hypotheses are assumed to be mutually exclusive, and thus compound events (where multiple outbreaks are occurring simultaneously) must be modeled as separate hypotheses. The posterior probability of each hypothesis (given the dataset D) is computed using Bayes' theorem:

$$Pr(H_1(S, E_k)|D) = \frac{Pr(D|H_1(S, E_k))Pr(H_1(S, E_k))}{Pr(D)}$$

$$Pr(H_0|D) = \frac{Pr(D|H_0)Pr(H_0)}{Pr(D)}$$

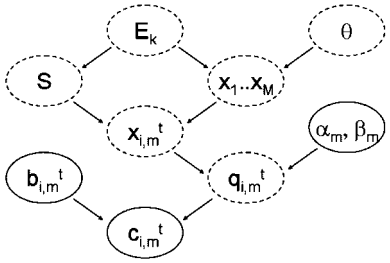

Figure 1. Bayesian network representation of the MBSS method. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. The counts $c_{i,m}^t$ are directly observed, whereas the baselines $b_{i,m}^t$ and the parameter priors for each stream (α_m, β_m) are estimated from historical data.

In this expression, the posterior probability of each hypothesis is normalized by the total probability of the data,

$$Pr(D) = Pr(D|H_0)Pr(H_0) + \sum_{S,E_k} Pr(D|H_1(S,E_k))Pr(H_1(S,E_k)).$$

The standard MBSS method [1] assumes a uniform prior $Pr(H_1(S,E_k))$ over all event types E_k and all circular spatial regions S. As discussed below, nonuniform priors can also be incorporated. In some cases, the prior distribution can be learned from a small number of labeled training examples [4].

The likelihood of the data given each hypothesis, $Pr(D|H)$, is computed using the hierarchical Gamma–Poisson model shown in Figure 1. Each count $c_{i,m}^t$ is assumed to be Poisson distributed: $c_{i,m}^t \sim \text{Poisson}(q_{i,m}^t b_{i,m}^t)$, where $b_{i,m}^t$ is the expected count (computed by time series analysis of historical data) and $q_{i,m}^t$ is the relative risk. Under the null hypothesis of no outbreaks, each relative risk $q_{i,m}^t$ is assumed to be generated from a Gamma distribution: $q_{i,m}^t \sim \text{Gamma}(\alpha_m, \beta_m)$, where α_m and β_m are estimated from the historical data for stream D_m, as described in [1]. Under the alternative hypothesis $H_1(S,E_k)$, the expected value of each relative risk is increased by a multiplicative factor $x_{i,m}^t$, the 'impact' of the outbreak for the given spatial location s_i, data stream D_m, and time step t: $q_{i,m}^t \sim \text{Gamma}(x_{i,m}^t \alpha_m, \beta_m)$. For a given set of impacts $X = \{x_{i,m}^t\}$, the likelihood of the data can be computed from the Gamma–Poisson model as follows [1]:

$$Pr(D|X) = \prod_{i,m,t} Pr(c_{i,m}^t | b_{i,m}^t, x_{i,m}^t, \alpha_m, \beta_m)$$

$$\propto \prod_{i,m,t} \left(\frac{\beta_m}{\beta_m + b_{i,m}^t}\right)^{x_{i,m}^t \alpha_m} \frac{\Gamma(x_{i,m}^t \alpha_m + c_{i,m}^t)}{\Gamma(x_{i,m}^t \alpha_m)}$$

In this expression, terms not dependent on the $x_{i,m}^t$ have been removed, since these are constant for all hypotheses under consideration. For the null hypothesis H_0, no events have occurred, and thus we assume $x_{i,m}^t = 1$ everywhere:

$$Pr(D|H_0) \propto \prod_{i,m,t} \left(\frac{\beta_m}{\beta_m + b_{i,m}^t}\right)^{\alpha_m} \frac{\Gamma(\alpha_m + c_{i,m}^t)}{\Gamma(\alpha_m)}$$

For the alternative hypothesis $H_1(S,E_k)$, we must marginalize over the values of $x_{i,m}^t$:

$$Pr(D|H_1(S,E_k)) = \sum_X Pr(D|X)Pr(X|H_1(S,E_k))$$

The distribution of the impacts $x_{i,m}^t$ is conditional on the outbreak type E_k, the affected region S, and two additional parameters: the outbreak severity θ and the temporal window W. The outbreak is assumed to affect those and only those locations in region S, for the most recent W time steps ($t = 0 \ldots W-1$). Thus we assume $x_{i,m}^t = 1$ for unaffected locations ($s_i \notin S$) and for time steps before the start of the outbreak ($t \geqslant W$). For affected locations and time steps ($s_i \in S$ and $t < W$), the impact $x_{i,m}^t$ is computed as a function of the 'average impact' $x_{\text{km,avg}}$ of outbreak type E_k on data stream D_m, and the outbreak severity θ:

$$x_{i,m}^t = 1 + \theta(x_{\text{km,avg}} - 1).$$

For example, if $x_{\text{km,avg}}$ was equal to 1.4, an outbreak of typical severity ($\theta = 1$) would increase counts by 40 per cent ($x_{i,m}^t = 1.4$), and an outbreak with severity $\theta = 2$ would increase counts by 80 per cent ($x_{i,m}^t = 1.8$).

The average effects of each outbreak type on each stream, $x_{km,avg}$, can either be specified by a domain expert or learned from training data, as described in [1]. For the experiments below, we assume a simple, fixed set of three outbreak models $E_1 \ldots E_3$ for two data streams D_1 and D_2. E_1 assumes that only stream D_1 is affected ($x_{11}=1.5$ and $x_{12}=1$), E_2 assumes that only stream D_2 is affected ($x_{21}=1$ and $x_{22}=1.5$), and E_3 assumes that the outbreak has equal impact on both data streams ($x_{31}=x_{32}=1.5$).

We assume a discrete uniform distribution Θ for θ, and assume that W is drawn uniformly between 1 and W_{max}, where W_{max} is the maximum temporal window size. W_{max} is an important parameter for detection: larger values of W_{max} improve detection power for more gradually emerging outbreaks, whereas smaller values of W_{max} improve detection power for more rapidly emerging outbreaks [5]. We consider two typical values, $W_{max}=3$ and $W_{max}=7$, in our experiments below. The total likelihood of the data given in the hypothesis $H_1(S, E_k)$ can be computed by marginalizing over θ and W:

$$Pr(D|H_1(S,E_k))=\frac{1}{W_{max}|\Theta|}\sum_{\theta\in\Theta}\sum_{W\in 1\ldots W_{max}} Pr(D|H_1(S,E_k),\theta,W)$$

As described in [1], the posterior probability map can be computed by MBSS in seven steps: loading the count data, computing baselines, computing parameter priors, computing location likelihood ratios, computing region likelihood ratios, computing region posterior probabilities, and computing the posterior probability map. The first three steps can be performed in time proportional to the size of the dataset, $O(NMT)$, where N, M, and T are the numbers of locations, streams, and time steps, respectively. The fourth step is to pre-compute the likelihood ratios

$$\mathrm{LR}_i=\prod_{m=1\ldots M}\prod_{t=0\ldots W-1}\frac{Pr(c^t_{i,m}|b^t_{i,m},x_m\alpha_m,\beta_m)}{Pr(c^t_{i,m}|b^t_{i,m},\alpha_m,\beta_m)}$$

for each location s_i, for each combination of the outbreak type E_k, temporal window W, and severity θ. Then the likelihood ratio for a given region S (given E_k, W, and θ) can be computed by multiplying the likelihood ratios LR_i for all $s_i \in S$. This formulation has the benefit that the expensive likelihood ratio computations are only performed a number of times proportional to the number of locations, rather than the much larger number of regions. The fifth step of the MBSS framework is to multiply the location likelihood ratios (or add log-likelihood ratios) to obtain the likelihood ratio for each spatial region S, for each combination of E_k, W, and θ. The sixth step requires computation of posterior probabilities for each combination of outbreak type E_k and spatial region S, marginalizing over W and θ, and the seventh step computes the posterior probability map by summing the posterior probabilities of all regions containing each spatial location. Each of the last three steps requires computation time proportional to the number of regions S, and thus MBSS is computationally expensive when the number of regions is large.

The standard MBSS method deals with this computational issue by considering only a small fraction of the $O(2^N)$ possible subsets of the N spatial locations, limiting its search to circular regions S and assuming that all non-circular regions have zero prior probability. As in Kulldorff's original spatial scan statistic [3], MBSS examines the set of $O(N^2)$ circular regions S, each containing a 'center' location s_c and its $k-1$ nearest neighbors (as measured by distance between the zip code centroids), for $k=1\ldots N$. As noted above, MBSS typically assumes a uniform region prior $Pr(H_1(S,E_k)|E_k)=1/N_S$, where N_S is the total number of space–time regions, but nonuniform priors can also be incorporated, and various methods for learning these priors from data [1, 4] have been explored.

2.2. *Fast subset sums*

As noted above, the primary limitation of the standard MBSS method is its exhaustive computation over spatial regions S, requiring us to restrict the set of spatial regions considered to only a small fraction of the $O(2^N)$ possible subsets of locations, e.g. only considering circular regions. This restriction prevents MBSS from searching over the huge number of irregularly shaped regions, reducing its detection power and spatial detection accuracy for highly elongated or irregular regions.

However, two key insights enable us to circumvent this limitation. First, both the total posterior probability of an outbreak and the posterior probability map are *sums* of the posterior region probabilities $Pr(H_1(S,E_k)|D)$, where the total posterior probability $Pr(H_1|D)$ is a sum over all spatial regions, and the posterior probability of each spatial location s_i is a sum over all spatial regions which contain s_i. We show that each of these sums can be calculated efficiently *without* computing the individual posterior probabilities of each spatial region S. Second, we define a specific, nonuniform prior distribution $Pr(H_1(S,E_k)|E_k)$ over all 2^N subsets of the data. This prior distribution has non-zero prior probabilities for any given subset of the data S, but more compact clusters have larger priors, thus enforcing a soft constraint on spatial proximity. Most importantly, as we demonstrate below, the prior distribution has a hierarchical structure which enables us to efficiently compute sums of posterior probabilities over exponentially many regions in linear time.

We consider a hierarchical region prior conditioned on two latent variables: the 'center' location s_c and the 'neighborhood size' k. We assume a two-stage process in which s_c and k are drawn from their respective distributions, and then the subset S is chosen conditional on s_c and k. As in the original MBSS method (searching over circular regions), we can assume that s_c is drawn uniformly at random from the set of spatial locations, and that the neighborhood size k is drawn uniformly at random between 1 and some constant $k_{\max}$ (the maximum neighborhood size). We typically use $k_{\max}=N$, the number of spatial locations, but smaller values of $k_{\max}$ can be used to reduce computation time, as shown below. Nonuniform distributions for the center s_c and neighborhood size k can easily be incorporated into our method, but the experiments described below assume uniform distributions. Once we have drawn the center location s_c and neighborhood size k, we then consider location s_c and its $k-1$ nearest neighbors, and choose a subset of locations *uniformly at random* from the 2^k possible subsets of these k locations. Thus the probability of a given center s_c, neighborhood k, and region S is $Pr(s_c,k,S)=Pr(s_c)Pr(k)/2^k$, if all locations $s_i\in S$ are contained in the k-neighborhood of s_c, and 0 otherwise. The total prior probability of a given region S can be computed by marginalizing over s_c and k. For uniform distributions of s_c and k, with $k_{\max}=N$, we can compute

$$Pr(S)=\frac{\sum_{s_c}2^{N-k_c+1}-1}{N^2 2^N},$$

where k_c is the minimum neighborhood size such that the k-neighborhood of s_c contains all elements of S. This expression is roughly proportional to $2^{-k_{\min}}$, where $k_{\min}=\min_{s_c}k_c$ is a measure of the compactness of region S.

The advantage of this formulation is that, for a given center location s_c and a given neighborhood size k, we can compute the total posterior probability of the 2^k spatial regions in $O(k)$ time. As we must consider $O(N)$ center locations and $O(k_{\max})$ neighborhood sizes, this approach enables us to compute the total posterior probability in time $O(Nk_{\max}^2)$. To do so, we define S_{ck} as the k-neighborhood of center location s_c (i.e. s_c and its $k-1$ nearest neighbors). Then, conditioning on the outbreak type E_k, severity θ, temporal window W, center location s_c, and neighborhood size k, we can compute the total posterior probability that any region $S\subseteq S_{ck}$ has been affected.

First, we know

$$\sum_{S\subseteq S_{ck}}Pr(S|D)\propto\sum_{S\subseteq S_{ck}}Pr(S)\mathrm{LR}(S),$$

where $Pr(S)$ is the prior probability of region S, conditioned on E_k, s_c, and k, and $\mathrm{LR}(S)$ is the likelihood ratio of region S, $Pr(D|H_1(S,E_k))/Pr(D|H_0)$, conditioned on E_k, θ, and W. Second, we know that

$$\sum_{S\subseteq S_{ck}}Pr(S)\mathrm{LR}(S)=\frac{1}{2^k}\sum_{S\subseteq S_{ck}}\mathrm{LR}(S),$$

as we are assuming a uniform prior over the 2^k subsets of S_{ck}. Third, we know that

$$\sum_{S\subseteq S_{ck}}\mathrm{LR}(S)=\sum_{S\subseteq S_{ck}}\prod_{s_i\in S}\mathrm{LR}_i,$$

where LR_i is the likelihood ratio of location s_i, conditioned on E_k, θ, and W. Fourth, as we are summing over all 2^k subsets of S_{ck}, we can write the sum of 2^k products as a product of k sums:

$$\sum_{S\subseteq S_{ck}}\prod_{s_i\in S}\mathrm{LR}_i=\prod_{s_i\in S_{ck}}(1+\mathrm{LR}_i).$$

Fifth, we obtain the final result,

$$\sum_{S\subseteq S_{ck}}Pr(S|D)\propto\frac{1}{2^k}\prod_{s_i\in S_{ck}}(1+\mathrm{LR}_i)=\prod_{s_i\in S_{ck}}\left(\frac{1+\mathrm{LR}_i}{2}\right).$$

Thus the posterior probability of an outbreak, conditioned on the outbreak type E_k, temporal window W, severity θ, center location s_c, and neighborhood size k, is proportional to the product of the smoothed likelihood ratios $(1+\mathrm{LR}_i)/2$ for all locations $s_i\in S_{ck}$. This is very similar to the case of circular regions, where the posterior probability of an outbreak (again, conditioned on E_k, W, θ, s_c, and k) was proportional to the product of the unsmoothed likelihood ratios LR_i. In each case, we must compute the total posterior probability of an outbreak by marginalizing over E_k, W, θ, s_c, and k.

Finally, we consider how the posterior probability map can be efficiently computed. To do so, we can efficiently compute the posterior probability of an outbreak affecting a given location s_j, conditioned on the outbreak type E_k, temporal window W, severity θ, center location s_c, and neighborhood size k, using a procedure very similar to the above. The only difference is that we must compute

$$\sum_{S\subseteq S_{ck}:s_j\in S}Pr(S|D)\propto\frac{1}{2^k}\sum_{S\subseteq S_{ck}:s_j\in S}\prod_{s_i\in S}\mathrm{LR}_i.$$

In this case, we are summing over all 2^{k-1} subsets of S_{ck} that contain s_j, and can write the sum of 2^{k-1} products as the product of $k-1$ sums:

$$\sum_{S\subseteq S_{ck}:s_j\in S}\prod_{s_i\in S}\mathrm{LR}_i=\mathrm{LR}_j\prod_{s_i\in S_{ck}-\{s_j\}}(1+\mathrm{LR}_i),$$

and thus

$$\sum_{S\subseteq S_{ck}:s_j\in S}Pr(S|D)=\left(\frac{\mathrm{LR}_j}{1+\mathrm{LR}_j}\right)\sum_{S\subseteq S_{ck}}Pr(S|D).$$

We can compute the total posterior probability of an outbreak containing each location s_j by marginalizing over E_k, W, θ, s_c, and k, thus enabling efficient computation of the posterior probability map.

2.3. Related work

This work describes FSS, an extension of our recently proposed MBSS framework [1] which enables computationally efficient detection of irregularly shaped outbreaks. MBSS extends our Bayesian event detection framework [6] to integrate information from multiple data streams, improving detection power and reducing false positives, and also enables us to accurately model and distinguish between multiple outbreak types. The Bayesian scan statistic is a variant of the more traditional, hypothesis testing approach to spatial scan statistics, developed by Kulldorff and Nagarwalla [3, 7]. While Kulldorff's original spatial scan statistic [3] did not take the time dimension into account, later work generalized this method to the 'space–time scan statistic' by scanning over variable size temporal windows [8, 9]. Recent extensions, such as the expectation-based scan statistic [10] and model-based scan statistic [11], use historical data to model the expected distribution of counts in each spatial location, and the expectation-based scan statistic approach is also applied in the present work to obtain the expected counts $b^t_{i,m}$.

Many other variants of the spatial and space–time scan statistics have been proposed, differing in both the set of regions to be searched and the underlying statistical models. Various statistical models have been proposed for the spatial scan, ranging from simple Poisson and Gaussian statistics [10, 12] to robust and nonparametric models [13, 14]. While Kulldorff's original method [3] assumed circular search regions, other methods have searched over rectangles [15], ellipses [16], and various sets of irregularly shaped regions [17–19]. We note that previous spatial scan methods for detecting irregularly shaped regions either restrict the search space (reducing power for any outbreak regions which do not fit the assumed distribution) or perform a heuristic search, in which case they are not guaranteed to find an optimal or near-optimal region; our FSS method, on the other hand, performs an exact computation over all $O(2^N)$ subsets of locations, though it computes the posterior probability map rather than identifying the specific subset of greatest interest.

Two multivariate extensions of the frequentist spatial scan have recently been proposed: Kulldorff's parametric scan [2], which directly extends the original spatial scan statistic to multiple data streams by assuming that all data streams are independent, and the nonparametric scan [14], which combines empirical p-values from multiple data streams without relying on an underlying parametric model. Unlike MBSS and FSS, neither of these two methods can differentiate between multiple outbreak types.

Several other multivariate disease surveillance methods have been proposed, including multivariate extensions of traditional time series analysis methods [20–22] and network-based methods [23] that detect anomalous ratios of counts between streams. These purely temporal methods do not take spatial information into account: they may be used to detect anomalous increases in the aggregate time series of the entire area being monitored, rather than detecting and pinpointing a spatial cluster of affected locations. Additionally, these methods cannot model and differentiate between multiple event types. PANDA [24, 25] uses Bayesian network models to differentiate between multiple outbreak types (e.g. the Centers for Disease Control (CDC) Category A diseases), assuming an underlying entity-based model of ED visits. We have recently developed a multivariate model that incorporates spatial information into PANDA, using ED chief complaint data as evidence [26].

We have recently developed other efficient search methods based on 'linear-time subset scanning' (LTSS), which allow us to efficiently perform unconstrained maximization of certain univariate likelihood ratio statistics over the 2^N subsets of the data [27]. These methods cannot be easily applied to the MBSS because of the added complications of non-uniform prior probabilities, multivariate data, and (most of all) the need to calculate the denominator of the posterior probability, which requires evaluation and summation of the posterior probabilities of all hypotheses under consideration. Additionally, LTSS methods do not guarantee a solution to constrained maximization problems, and thus are difficult to apply in spatial detection scenarios where we wish to enforce spatial proximity constraints on the detected regions.

3. Evaluation

We now compare the run time, detection power, and spatial accuracy of the FSS method to the original MBSS. While FSS and MBSS both rely on the Bayesian probabilistic framework [1] described briefly above, MBSS assumes a uniform prior over the set of $O(N^2)$ circular regions, whereas FSS assumes a nonuniform, hierarchical prior which is nonzero for all $O(2^N)$ subsets of the data. Thus we expect FSS to improve timeliness and accuracy of detection for elongated or irregular outbreak regions, as these are not modeled well by the assumption of circular regions. However, since a much larger set of regions must be considered, we expect the run time of FSS to be slower than the original MBSS method. However, we expect the run time of FSS to be much faster than a naive search over all $O(2^N)$ subsets, which would be computationally infeasible for even moderate values of N. The following subsections describe our test dataset (ED data from Allegheny County), compare the run time of FSS to the MBSS and naive subset sums methods, describe our outbreak simulations, and compare the detection power and spatial accuracy of FSS and MBSS for these simulated outbreaks.

3.1. Description of emergency department data

We obtained a dataset of 612 713 de-identified ED visit records collected from 10 Allegheny County hospitals from January 1, 2004 to December 31, 2005. Each record contains fields for the patient's date of admission to the ED, home zip code, chief complaint (free text), and International Classification of Diseases (ICD9) code (numeric). We removed records where the home zip code or admission date was missing, or where the home zip code was outside Allegheny County, leaving 397 134 records (64.8 per cent). The free-text chief complaint was present for all remaining records, and the ICD9 code was present for 336 338 (84.7 per cent) of the remaining records.

From this data, we created two distinct streams of count data (cough/dyspnea (CD) and nausea/vomiting (NV)) by recording the number of patient records matching the given symptoms in each zip code for each day. A patient record was determined to match a given set of symptoms if its chief complaint string contained certain substrings, or if its ICD9 code matched certain values: for the CD stream, we included records with chief complaints containing the substrings 'cough', 'dyspnea', 'shortness', or 'sob', or with ICD9 codes matching 786.2 (cough) or 786.05 (shortness of breath). For the NV stream, we included records with chief complaints containing the substrings 'naus', 'vom', or 'n/v', or with ICD9 codes matching 787.01–787.03 (nausea and/or vomiting) or 536.2 (persistent vomiting). Each set of records was manually refined to remove spurious substring matches.

The time series of daily counts (aggregated over all 97 Allegheny County zip codes) for each data stream is shown in Figure 2. The CD stream had a mean daily count of 44.0 cases, with a standard deviation of 12.1, and the NV stream had a mean daily count of 25.9 cases, with a standard deviation of 7.0. As these cases were spread over the 97 Allegheny County zip codes, many zip codes had zero counts on any given day. Both streams were weakly overdispersed, with index of dispersion (ratio of variance to mean) 3.31 for the CD data and 1.88 for the NV data, and daily counts of the two streams were positively correlated ($r = 0.338$). Both streams exhibited slight (but statistically significant) day-of-week trends, with counts peaking on Mondays, and clear seasonal trends, with counts peaking in February and March for the CD and NV datasets, respectively.

3.2. Comparison of run times

Our first experiment compared the run times of FSS, Naive Subset Sums (searching exhaustively over all subsets of locations), and the original MBSS method (searching over circular regions), as a function of the maximum neighborhood

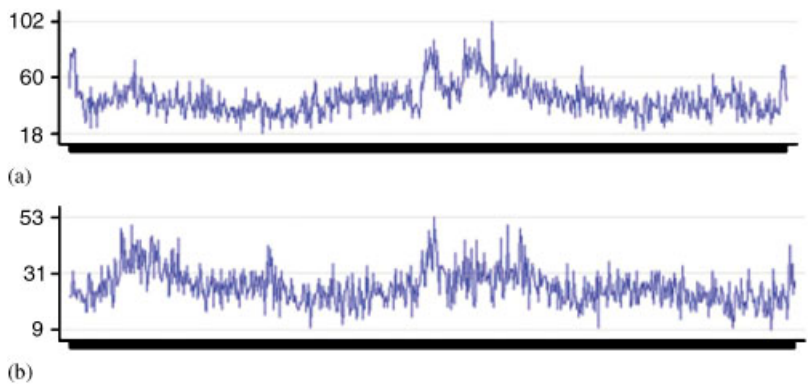

Figure 2. Daily counts for two streams of Allegheny County Emergency Department data (cough/dyspnea and nausea/vomiting) from January 1, 2004 to December 31, 2005: (a) CD and (b) NV.

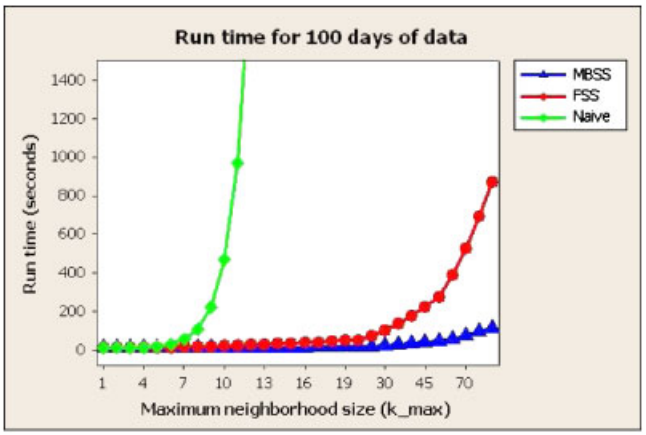

Figure 3. Total run times for 100 days of data, for Fast Subset Sums (FSS), naive subset sums, and the original MBSS approach, as a function of the maximum neighborhood size $k_{\max}$.

size $k_{\max}$. Increasing the neighborhood size requires a larger number of regions to be searched: the number of search regions with neighborhood size $k \in \{1 \ldots k_{\max}\}$ scales linearly with $k_{\max}$ if only circular regions are considered, and exponentially with $k_{\max}$ if we consider all subsets of the center and its $k-1$ nearest neighbors.

We computed the time required by each method to compute the posterior probability maps for the first 100 days of Allegheny County ED data, for each value of $k_{\max}$. For each method, we used the same two data streams (CD and NV), the same three outbreak models (assuming one outbreak type that affects only CD, one outbreak type that affects only NV, and one that affects both streams equally), and the same maximum temporal window size of three days ($W_{\max}=3$).

As shown in Figure 3, the run time of the original MBSS method increased gradually with increasing $k_{\max}$, up to a maximum of 116.90 s (to process 100 days of data) for $k_{\max}=90$. The run time of the Naive Subset Sums method increased exponentially, making it computationally infeasible for $k_{\max} \geqslant 25$. Run times were 5.02 h for 100 days of data for $k_{\max}=15$, and 161 h for 100 days of data for $k_{\max}=20$; for $k_{\max}=30$, the naive method would have required an estimated 68.5 days to process a single day of data, and for $k_{\max}=90$, it would have required an estimated 2×10^{17} years. Finally, we observe that the run time of FSS scales quadratically with increasing $k_{\max}$, up to a maximum of 876.25 s (to process 100 days of data) for $k_{\max}=90$. Thus, while FSS is approximately 7.5× slower than the original MBSS method, it is still extremely fast, computing the posterior probability map for each day of data in under 9 s.

We also considered a fixed $k_{\max}=10$, and examined the effects of doubling the number of data streams (from 2 to 4), doubling the number of event models (from 3 to 6), and doubling the maximum temporal window size (from 3 to 6), respectively. Each of these three changes increased the run time of each method by 30–80 per cent, demonstrating linear dependence of run time on these three parameters as expected.

3.3. Simulation of outbreaks

Next we used a semi-synthetic testing framework (injecting simulated multivariate outbreaks into the real-world ED data) to compare the detection power and spatial accuracy of the FSS and MBSS methods. We considered a simple class of simulated outbreaks with a linear increase in the expected number of cases over the duration of the outbreak. More precisely, our outbreak simulator takes three parameters: the outbreak duration T, the outbreak severity Δ, and the subset of affected zip codes S_{inject}. Then for each injected outbreak, the outbreak simulator chooses the start date of the outbreak t_{start} uniformly at random. On each day t of the outbreak, $t=1 \ldots T$, the outbreak simulator injects Poisson($t w_{i,m} \Delta_m$) cases into each stream of each affected zip code, where $w_{i,m}$ is the 'weight' of the zip code for that stream,

$$w_{i,m} = \frac{\sum_t c_{i,m}^t}{\sum_i \sum_t c_{i,m}^t}.$$

We considered 10 differently shaped outbreak regions S_{inject}, as shown in Figures 5–14 (left panels). All outbreaks were assumed to be two weeks in duration ($T=14$), and we assumed $\Delta_{\text{CD}}=\Delta_{\text{NV}}=1$. For each outbreak region, we considered 200 different, randomly generated outbreaks, giving a total of 2000 outbreaks for evaluation.

We note that simulation of outbreaks is an active area of ongoing research in biosurveillance. The creation of realistic outbreak scenarios is important because of the difficulty of obtaining sufficient labeled data from real outbreaks, but is also very challenging. State-of-the-art outbreak simulations, such as those of Buckeridge *et al.* [28], and Wallstrom

Table I. Comparison of MBSS and FSS methods, using maximum temporal window size $W_{\max} = 3$.

Outbreak region	MBSS	FSS
1	8.275 (96.0 per cent)	7.550 (98.0 per cent)
2	8.045 (96.0 per cent)	7.480 (96.5 per cent)
3	10.275 (71.5 per cent)	9.095 (85.0 per cent)
4	9.780 (83.5 per cent)	8.880 (90.5 per cent)
5	10.425 (68.5 per cent)	8.310 (95.0 per cent)
6	7.865 (97.5 per cent)	6.810 (99.5 per cent)
7	7.400 (100 per cent)	6.515 (100 per cent)
8	4.230 (100 per cent)	3.975 (100 per cent)
9	5.900 (100 per cent)	5.205 (100 per cent)
10	7.440 (99.0 per cent)	5.930 (100 per cent)
Average	7.964 (91.2 per cent)	6.975 (96.5 per cent)

Average days to detection, and proportion of outbreaks detected, at a fixed false-positive rate of 1/month.

Table II. Comparison of MBSS and FSS methods, using maximum temporal window size $W_{\max} = 7$.

Outbreak region	MBSS	FSS
1	8.860 (97.0 per cent)	8.995 (97.0 per cent)
2	8.200 (99.0 per cent)	8.265 (99.0 per cent)
3	11.380 (65.0 per cent)	10.835 (79.0 per cent)
4	10.530 (78.5 per cent)	10.195 (83.5 per cent)
5	11.665 (62.5 per cent)	9.470 (94.5 per cent)
6	8.310 (99.0 per cent)	7.845 (100 per cent)
7	7.555 (100 per cent)	7.245 (100 per cent)
8	4.700 (100 per cent)	4.610 (100 per cent)
9	6.295 (100 per cent)	5.920 (100 per cent)
10	7.600 (100 per cent)	6.340 (100 per cent)
Average	8.510 (90.1 per cent)	7.972 (95.3 per cent)

Average days to detection, and proportion of outbreaks detected, at a fixed false-positive rate of 1/month.

et al. [29], combine disease trends observed from past outbreaks with information about the current background data into which the outbreak is being injected, as well as allowing the user to adjust parameters, such as outbreak duration and severity. While the simple linear outbreak model that we use here is not a realistic model of the temporal progression of an outbreak, it enables precise comparison of the detection power of different methods, gradually ramping up the severity of the outbreak until it is detected.

3.4. Comparison of detection power

We compared the detection power of the FSS and MBSS methods for each of the 10 outbreak regions, considering two values of the maximum temporal window size ($W_{\max} = 3$ and $W_{\max} = 7$) for each method. We fixed the value of $k_{\max} = N$, i.e. the MBSS method considers all $O(N^2)$ circular regions and the FSS method considers all $O(2^N)$ subsets of the data. For each combination of method and outbreak region, we computed the method's proportion of outbreaks detected and average number of days to detect 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 (the first 84 days of data are excluded, since these are used to calculate baseline estimates for our methods). Then for each injected outbreak, we computed the maximum region score for each outbreak day, and determined what proportion of the days for the original dataset have 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 ($t = 1 \ldots 14$) with proportion of false positives less than r. If no day of the outbreak has 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 counted as 14 days to detect, but could also be penalized further.

The detection performance of the MBSS and FSS methods is presented in Tables I and II. For each of the 10 outbreak regions, we present each method's average detection time and percentage of outbreaks detected at a fixed false-positive rate of 1/month. We also show the AMOC curves (average detection time as a function of false-positive rate) for MBSS and FSS, averaged over all 10 outbreak regions, in Figure 4.

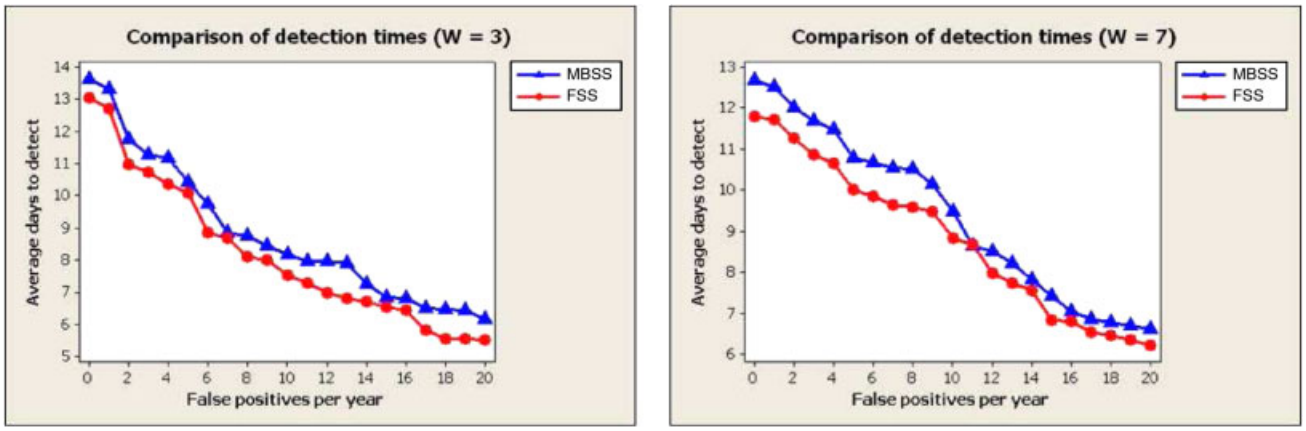

Figure 4. AMOC curves for FSS and MBSS, for maximum temporal window sizes 3 and 7. Average days to detection as a function of false-positive rate.

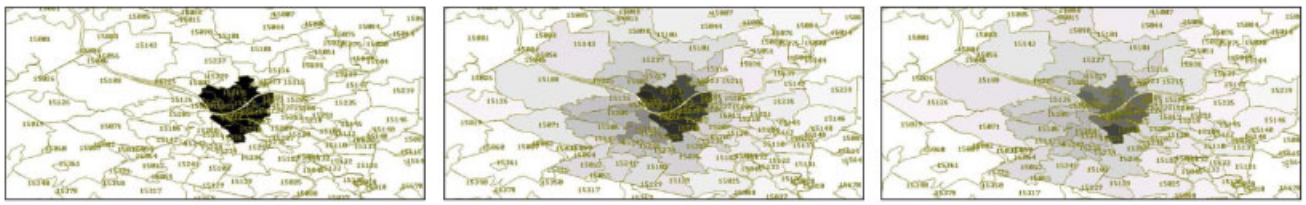

Figure 5. Outbreak region #1 (compact cluster, downtown Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

From the AMOC curves, we see that FSS consistently achieves more timely detection than MBSS across a range of false-positive rates. For a fixed false-positive rate of 1/month, FSS detected an average of one day earlier than MBSS for a maximum temporal window size $W_{\max}=3$, and 0.54 days earlier for $W_{\max}=7$, with fewer than half as many missed outbreaks in each case. Comparing the detection performance of FSS and MBSS for each of the 10 outbreak regions, we observe that both methods achieve very similar detection times for compact outbreak regions (#1, #2, and #8). For highly elongated outbreak regions (#5 and #10), FSS detected between 1.3 and 2.2 days earlier. For moderately elongated and irregularly shaped regions, the difference between methods was smaller: FSS detected 0.7–1.2 days earlier for $W_{\max}=3$ and 0.3–0.6 days earlier for $W_{\max}=7$. Both methods achieved more timely detection when $W_{\max}=3$, suggesting that the given outbreaks did not emerge gradually enough for the increased temporal window size to improve performance.

3.5. Comparison of spatial accuracy

For each of the 10 outbreak regions, we compared the spatial accuracy of the MBSS and FSS methods. The center panels of Figures 5–14 show the average posterior probability maps for the original MBSS method for each of the 10 outbreak regions at the midpoint of the outbreak, averaged over the 200 randomly generated outbreaks affecting that region. The right panels of Figures 5–14 show the average posterior probability maps for FSS for each outbreak region at the midpoint of the outbreak. We used a maximum temporal window size $W_{\max}=7$ for both methods. Comparing each averaged map to the true affected regions (left panels of Figures 5–14), we observe that FSS more accurately estimates the spatial extent of the outbreak region. The detected regions were very similar for the compact outbreaks (#1, #2, and #8). For the elongated outbreaks (#4, #5, #9, and #10), FSS closely approximated the true outbreak region, whereas MBSS was not able to accurately distinguish the true region. Finally, for the irregular and disjoint outbreaks (#3, #6, and #7), FSS was better able to fit the irregular shape, whereas MBSS typically reported a compact cluster containing the irregular shape and several additional (incorrect) zip codes.

To quantify these differences in spatial detection accuracy, we computed the average overlap coefficient, precision, and recall for each method for each of the 10 outbreak types, as a function of the outbreak day. To compute the overlap coefficient for a given day of a given outbreak, we first compute the set of 'detected' zip codes. To do so, we find the zip code with largest posterior outbreak probability p, and consider any zip code with posterior probability greater than $p/2$ to have been detected. We compare the 'detected' zip codes to the set of 'true' zip codes containing injected cases, and count the number of 'hits' (zip codes which are both true and detected). The overlap coefficient is then defined as

Figure 6. Outbreak region #2 (compact cluster, southeast of Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

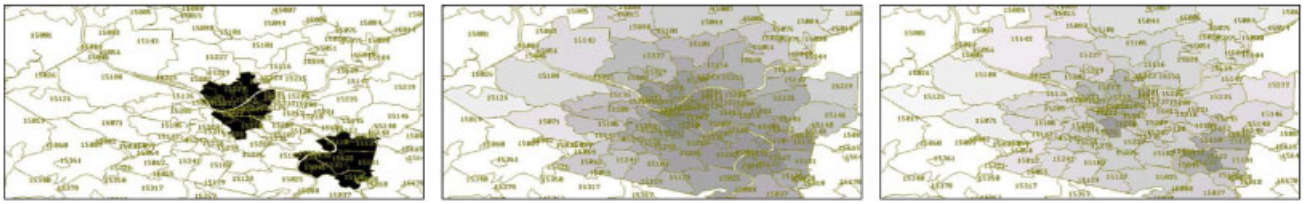

Figure 7. Outbreak region #3 (two disjoint clusters). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

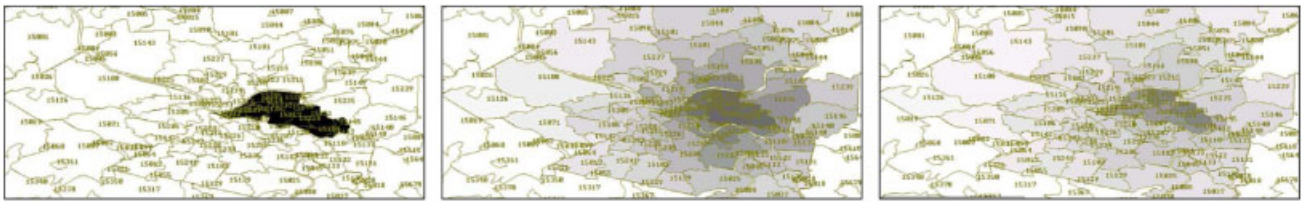

Figure 8. Outbreak region #4 (elongated cluster, downtown Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

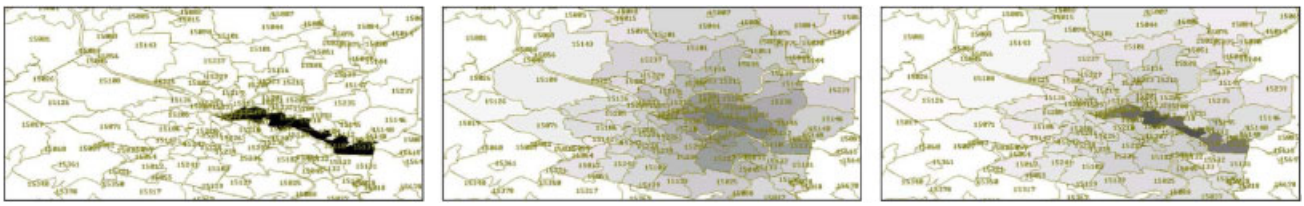

Figure 9. Outbreak region #5 (highly elongated cluster, downtown Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

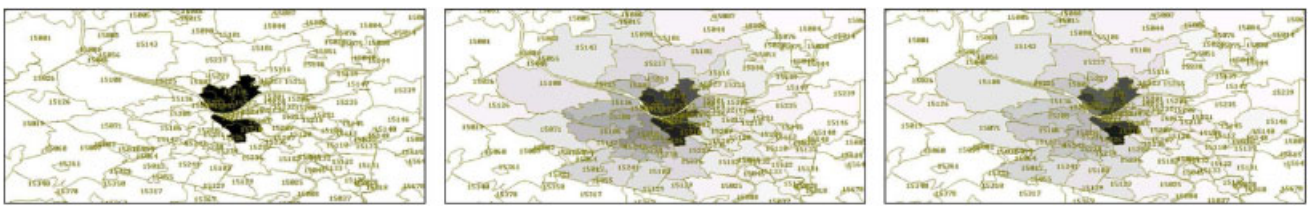

Figure 10. Outbreak region #6 (irregular cluster, Pittsburgh North and South Sides). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

$N_{\text{hits}}/(N_{\text{true}} + N_{\text{detected}} - N_{\text{hits}})$. The precision is defined as $N_{\text{hits}}/N_{\text{detected}}$, and the recall is defined as $N_{\text{hits}}/N_{\text{true}}$. Finally, we compute the average overlap coefficient, precision, and recall for each outbreak region (#1–10) for each outbreak day (1–14), averaging over all 200 outbreaks of the given type.

Figure 11. Outbreak region #7 (irregular cluster, Monroeville area). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

Figure 12. Outbreak region #8 (compact cluster, west of Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

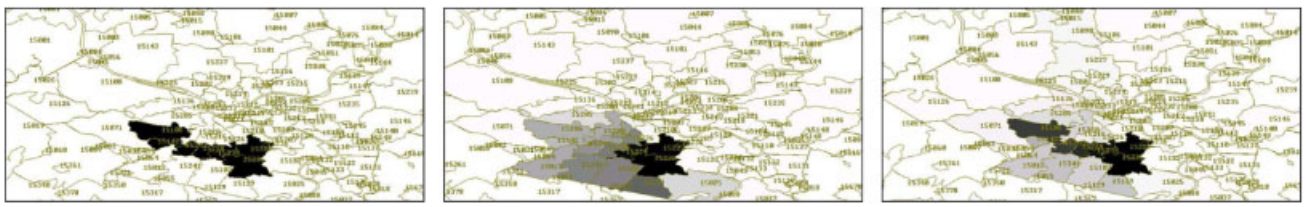

Figure 13. Outbreak region #9 (elongated cluster, south of Pittsburgh). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

Figure 14. Outbreak region #10 (highly elongated cluster, north Allegheny County). Left: true outbreak region. Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

Figure 15 shows the average overlap coefficients (averaged over all 10 outbreak regions) for each method as a function of the outbreak day. Tables III and IV show the average overlap coefficient, precision, and recall for each outbreak region for each method, at the midpoint of the outbreak. From the figure, we observe that FSS increased the average overlap coefficient by 6–15 per cent when compared with MBSS, for a given outbreak day and a given value of the maximum temporal window size $W_{\max}$. Both methods had higher overlap coefficients for $W_{\max}=7$: as many of the affected locations may have had no counts injected on a given time step, a longer temporal window enabled the methods to observe a larger number of injected counts, thus improving their ability to pinpoint the affected region.

At the midpoint of the outbreak, FSS increased the average overlap coefficient by 7.3 per cent for $W_{\max}=3$ and 10.0 per cent for $W_{\max}=7$. However, we observe from the tables that the overlap coefficients were similar for the more compact outbreaks (#1, #2, and #7), while FSS substantially improved the overlap coefficients for the more elongated outbreaks (#5, #9, and #10). FSS obtained substantially higher precision than MBSS for all 10 outbreak regions, whereas MBSS obtained substantially higher recall for 8 of 10 outbreak regions. Increasing $W_{\max}$ from 3 to 7 improved both precision and recall for FSS, whereas only improving precision for MBSS. FSS tended to miss affected zip codes with very small numbers of injected counts, whereas MBSS tended to include these zip codes if they were spatially proximate

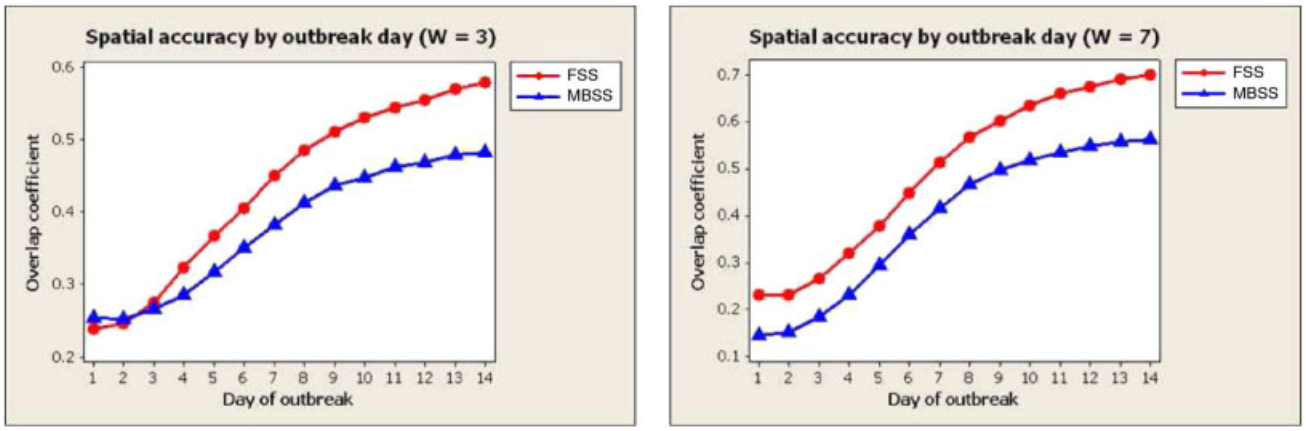

Figure 15. Spatial detection accuracy for FSS and MBSS, for maximum temporal window sizes 3 and 7. Average overlap coefficient between true and estimated regions for each outbreak day.

Table III. Comparison of MBSS and FSS methods, using maximum temporal window size $W_{\max} = 3$.

	Precision		Recall		Overlap	
Outbreak region	MBSS (per cent)	FSS (per cent)	MBSS (per cent)	FSS (per cent)	MBSS (per cent)	FSS (per cent)
1	51.4	67.0	95.4	71.4	49.7	50.1
2	62.6	66.3	91.1	64.8	57.9	44.9
3	36.5	51.9	79.2	53.6	29.6	31.3
4	40.4	52.2	90.5	70.1	35.7	37.0
5	18.4	45.0	84.6	73.2	14.9	35.8
6	40.4	59.0	93.4	76.5	38.6	48.8
7	65.9	78.1	98.2	81.4	65.5	65.8
8	57.3	82.2	93.8	81.1	55.1	68.3
9	61.2	80.6	71.0	76.4	40.7	62.9
10	37.5	64.2	60.4	59.8	24.9	40.9
Average	47.2	64.7	85.8	70.8	41.3	48.6

Average overlap coefficient, precision, and recall at the midpoint of the outbreak.

Table IV. Comparison of MBSS and FSS methods, using maximum temporal window size $W_{\max} = 7$.

	Precision		Recall		Overlap	
Outbreak region	MBSS (per cent)	FSS (per cent)	MBSS (per cent)	FSS (per cent)	MBSS (per cent)	FSS (per cent)
1	66.0	73.0	97.4	76.5	64.3	58.5
2	68.7	80.3	90.4	65.9	63.2	54.7
3	33.9	50.8	84.6	56.2	28.8	34.5
4	40.9	58.9	89.8	68.8	35.9	42.9
5	24.1	53.1	82.7	75.6	18.7	41.4
6	50.0	66.0	95.1	83.4	48.5	58.0
7	77.8	90.9	99.3	86.3	77.7	79.2
8	61.2	85.2	94.2	86.3	58.8	74.3
9	66.9	83.6	65.8	84.0	41.4	71.1
10	44.1	69.5	56.9	68.4	29.0	51.3
Average	53.4	71.1	85.6	75.1	46.6	56.6

Average overlap coefficient, precision, and recall at the midpoint of the outbreak.

to other affected zip codes; on the other hand, MBSS often detected additional spatially proximate zip codes which were not part of the affected region. The most likely reason for these differences is that FSS picks out the (possibly irregularly shaped) subset of locations with sufficiently many injected cases without including the surrounding locations, whereas MBSS typically detects a larger, circular region containing these locations. When the true outbreak region is compact (circular or nearly circular), the circular region chosen by MBSS may better approximate its true extent; for elongated

or irregular regions, the circular region chosen by MBSS is a poor approximation, and the greater flexibility of FSS is needed to improve spatial accuracy.

4. Conclusions

The FSS method is an extension of our MBSS framework [1] which allows the computationally efficient detection of irregular clusters. By defining a hierarchical prior over all $O(2^N)$ subsets of the N monitored locations, rather than considering only the $O(N^2)$ circular regions, FSS can achieve more timely detection of elongated or irregularly shaped clusters, and can more accurately pinpoint the affected subset of locations. Our experiments demonstrate substantial improvements in detection time (1.3–2.2 days) and spatial accuracy (more than doubling the overlap coefficient between true and detected regions) for highly elongated clusters, while achieving similar detection time and spatial accuracy for compact clusters. Most importantly, while a naive approach to computing the posterior probability map (requiring likelihood computations for each of the exponentially many subsets of the data) would be computationally infeasible, our computationally efficient FSS method scales up to the approximately 100 zip codes in Allegheny County, requiring less than 9 s to compute the posterior probability map for a given day of data.

Our continuing exploration of the FSS approach will extend the present work in two main directions. First, while our evaluation of detection power and spatial accuracy assumed a maximum neighborhood size $k_{\max}$ equal to the number of locations N, $k_{\max}$ can also be set to less than N if we wish to rule out highly elongated clusters, improving detection power for compact and nearly compact clusters, and reducing computation time. Similarly, while we assumed that each subset of the chosen neighborhood (a center location s_c and its $k-1$ nearest neighbors) was equally likely, we can also generalize to the case where we independently choose (with some probability p) whether each location in the neighborhood is affected. Our current method is equivalent to assuming $p=\frac{1}{2}$, and searching over circles is equivalent to assuming $p=1$, but we can also choose a value of p between $\frac{1}{2}$ and 1. This generalization still allows us to consider elongated and irregular regions, but gives a higher weight to more compact regions. Our future work will perform a detailed analysis of the detection power of the generalized FSS method as a function of $k_{\max}$, p, and the compactness of the outbreak region.

Finally, we propose to incorporate incremental model learning into the FSS approach. As in our original MBSS method, the average effects of each outbreak type on each data stream can be learned from a small number of labeled training examples using a smoothed maximum likelihood approach [1]. Learning the hierarchical prior over subsets of the data is more challenging, as we must infer the distributions over center locations s_c and neighborhood size k (and optionally, the location probability p) from partially labeled training examples. Typically, the subset of affected locations S is labeled, but the values of s_c, k, and p are unknown. We will apply a generalized expectation-maximization (GEM) approach to estimate the distributions of these values, assuming multinomial distributions for s_c and k, and a fixed (but unknown) parameter p. This model is very similar to the 'latent center' model described in [4], but conditions on the neighborhood size k rather than the radius r, and assumes that the conditional probability that a location is affected is uniform (within the chosen neighborhood) rather than decreasing as a function of distance from the center. We believe that incorporation of learning into the FSS approach will enable efficient optimization of the hierarchical prior from a small number of training examples, further improving the timeliness and accuracy of outbreak detection.

Acknowledgements

This work was partially supported by the National Science Foundation grants IIS-0325581, IIS-0916345, and IIS-0911032.

References

1. Neill DB, Cooper GF. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 2009; DOI: 10.1007/s10994-009-5144-4.
2. Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. *Statistics in Medicine* 2007; **26**:1824–1833.
3. Kulldorff M. A spatial scan statistic. *Communications in Statistics*: *Theory and Methods* 1997; **26**(6):1481–1496.
4. Makatchev M, Neill DB. Learning outbreak regions in Bayesian spatial scan statistics. *Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health Care Applications*, Helsinki, Finland, 2008.
5. Neill DB. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting* 2009; **25**:498–517.
6. Neill DB, Moore AW, Cooper GF. A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems 18*, 2006; 1003–1010.
7. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810.

8. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: a space–time scan statistic and cluster alarms in Los Alamos. *American Journal of Public Health* 1998; **88**:1377–1380.
9. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A* 2001; **164**:61–72.
10. Neill DB, Moore AW, Sabhnani MR, Daniel K. Detection of emerging space–time clusters. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, U.S.A., 2005.
11. Kleinman K, Abrams A, Kulldorff M, Platt R. A model-adjusted space–time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection* 2005; **133**(3):409–419.
12. Neill DB. Detection of spatial and spatio-temporal clusters. *Technical Report CMU-CS-06-142*, *Ph.D. Thesis*, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA, 2006.
13. Neill DB, Sabhnani MR. A robust expectation-based spatial scan statistic. *Advances in Disease Surveillance* 2007; **2**:61.
14. Neill DB, Lingwall J. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 2007; **4**:106.
15. Neill DB, Moore AW, Sabhnani MR. Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report* 2005; **54**(Supplement on Syndromic Surveillance):197.
16. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; **25**:3929–3943.
17. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis* 2004; **45**:269–286.
18. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 2004; **11**:183–197.
19. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.
20. Burkom HS. Biosurveillance applying scan statistics with multiple, disparate data sources. *Journal of Urban Health* 2003; **80**(2 Suppl. 1):i57–i65.
21. Burkom HS, Murphy SP, Coberly J, Hurt-Mullen K. Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report* 2005; **54**(Supplement on Syndromic Surveillance):55–62.
22. Mohtashemi M, Kleinman K, Yih WK. Multi-syndrome analysis of time series using PCA: a new concept for outbreak investigation. *Statistics in Medicine* 2007; **26**(29):5203–5224.
23. Reis BY, Kohane IS, Mandl KD. An epidemiological network model for disease outbreak detection. *PLoS Medicine* 2007; **4**:210.
24. Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM. Bayesian biosurveillance of disease outbreaks. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Banff, Canada, 2004.
25. Cooper GF, Dowling JN, Levander JD, Sutovsky P. A Bayesian algorithm for detecting CDC Category A outbreak diseases from emergency department chief complaints. *Advances in Disease Surveillance* 2007; **2**:45.
26. Jiang X, Neill DB, Cooper GF. A Bayesian network model for spatial event surveillance. *International Journal of Approximate Reasoning* 2010; **51**:224–239.
27. Neill DB. Fast and flexible outbreak detection by linear-time subset scanning. *Advances in Disease Surveillance* 2008; **5**:48.
28. Buckeridge DL, Burkom HS, Moore AW, Pavlin JA, Cutchis PN, Hogan WR. Evaluation of syndromic surveillance systems: development of an epidemic simulation model. *Morbidity and Mortality Weekly Report* 2004; **53**(Supplement on Syndromic Surveillance):137–143.
29. Wallstrom GL, Wagner MM, Hogan WR. High-fidelity injection detectability experiments: a tool for evaluation of syndromic surveillance systems. *Morbidity and Mortality Weekly Report* 2005; **54**(Supplement on Syndromic Surveillance):85–91.