Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 2, *pp.* 337–360

Fast subset scan for spatial pattern detection

Daniel B. Neill

Carnegie Mellon University, Pittsburgh, USA

[Received July 2011. Final revision August 2011]

Summary. We propose a new 'fast subset scan' approach for accurate and computationally efficient event detection in massive data sets. We treat event detection as a search over subsets of data records, finding the subset which maximizes some score function. We prove that many commonly used functions (e.g. Kulldorff's spatial scan statistic and extensions) satisfy the 'linear time subset scanning' property, enabling exact and efficient optimization over subsets. In the spatial setting, we demonstrate that proximity-constrained subset scans substantially improve the timeliness and accuracy of event detection, detecting emerging outbreaks of disease 2 days faster than existing methods.

Keywords: Algorithms; Disease surveillance; Event detection; Scan statistics; Spatial scan

1. Introduction

This work develops new methods for accurate and computationally efficient detection of emerging events in massive spatial and space–time data sets. Event detection is a ubiquitous task with a wide variety of real world applications: for example, agencies that are responsible for public health and safety must respond rapidly to potential threats including outbreaks of disease, terrorist attacks and natural disasters. In the event detection task, we must identify whether there are any interesting or anomalous patterns in the data and characterize each pattern by pinpointing the subset of data records affected. For example, in disease surveillance, we wish to identify whether there are any emerging outbreaks of disease, which areas have been affected and how long the outbreak has been going on. Such events must be detected in the very early stages, requiring identification of subtle patterns (e.g. a 20% increase in cases of fever at three local hospitals) in noisy background data. These subtle signals may not be detectable if we examine only a small part of the subset affected (a single hospital) or a larger subset containing many unaffected records (the aggregate count for the entire city). As a result, both 'bottom-up' approaches which identify and aggregate individual anomalous records (Barnett and Lewis, 1994) and 'top-down' approaches which detect anomalous global trends in the data often have low power to detect emerging events.

This suggests an alternative *subset scan* approach, where we search over subsets of the data (e.g. groups of data records) and identify those subsets that correspond to potentially relevant patterns. More precisely, we can define a score function $F(S)$ which measures the 'interestingness' or 'anomalousness' of a subset S and perform constrained or unconstrained maximization of $F(S)$ over all subsets of the data. This very general formulation requires us to define an appropriate score function $F(S)$ and to address the computational challenge of maximizing $F(S)$ over subsets of the data. For spatial and space–time data, a variety of 'spatial scan statistics' have

Address for correspondence: Daniel B. Neill, H. J. Heinz III College, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.
E-mail: neill@cs.cmu.edu

 1369–7412/12/74337

been developed. These methods, which are discussed in detail below, maximize a likelihood ratio statistic over spatial regions (subsets of locations). Since the number of possible subsets scales exponentially with the size of the data set, an exhaustive search over subsets is computationally infeasible. Typical spatial scan methods either restrict the search space or perform an approximate heuristic search, resulting in reduced power of detection and lower accuracy. Here we propose an alternative approach: new *fast subset scan* methods that efficiently identify the most anomalous subsets without an exhaustive search.

We demonstrate that many useful score functions $F(S)$ satisfy a property ('linear time subset scanning' (LTSS)) which allows extremely efficient *unconstrained* optimization over all subsets of the data. For score functions satisfying LTSS, the subset of data records which maximizes $F(S)$ can be found by ordering the records according to some 'priority' function and searching over groups consisting of the top k highest priority records, requiring only a linear rather than exponential number of subsets to be evaluated. In the spatial setting, we demonstrate that many commonly used spatial scan statistics, including Kulldorff's original spatial scan statistic (Kulldorff, 1997) and many recently proposed variants such as the expectation-based Poisson (Neill *et al.*, 2005), Gaussian (Neill, 2006) and exponential (Huang *et al.*, 2007) scan statistics, satisfy the LTSS property, allowing us to find the most anomalous subset of N locations while searching only N subsets rather than 2^N. However, we often wish to incorporate *spatial proximity* constraints, thus identifying a spatial region (a group of nearby locations) rather than a spatially dispersed set of locations. We show that, for score functions satisfying LTSS, efficient spatially constrained optimization can be performed by using multiple unconstrained optimization steps, thus enabling rapid identification of the most interesting spatial regions.

In certain cases, the unconstrained fast subset scan approach reduces to a variant of the upper level set (ULS) scan statistic that was proposed by Patil and Taillie (2004), but the ULS also enforces connectivity constraints on the cluster detected. However, our method is more general than the ULS: it is applicable for optimization of a large class of score functions, can incorporate a variety of constraints and can be extended to non-spatial and multivariate data. Moreover, a simple counterexample (Section 3.6) shows that the ULS is not guaranteed to compute the highest scoring connected cluster, whereas we prove that fast subset scan can efficiently find the exact solution to constrained and unconstrained subset scan problems without an exhaustive search.

2. Spatial event detection

Although our proposed fast subset scan framework has the potential to be applied to many different types of data, here we focus on the problem of *spatial event detection*, in which we monitor spatial time series data with the goal of rapidly detecting and identifying emerging patterns. For example, in *spatial disease surveillance*, we monitor electronically available public health data such as hospital visits and medication sales to detect emerging outbreaks of disease. Early and accurate detection of outbreaks is of critical importance: major health threats such as emerging infectious diseases or bioterrorist attacks require rapid and appropriate responses to control the spread of disease, treat infected individuals and reduce the potentially catastrophic costs to society.

In the spatial event detection problem, we monitor a set of data streams $\{D_1 \ldots D_M\}$ over time at a set of spatial locations $\{s_1 \ldots s_N\}$. For each stream D_m and location s_i, we are given a time series of observed real-valued counts $c^t_{i,m}$. For example, in disease surveillance, each data stream could represent the number of hospital visits corresponding to a different category of symptom (respiratory, fever, etc.) For data collected daily and aggregated at the zip code level,

a given count $c^t_{i,m}$ might represent the number of respiratory cases for zip code s_i on day t. For each data stream D_m and location s_i, we first compute the time series of expected counts (or 'baselines') $b^t_{i,m}$ by using the historical data for that stream and location (Neill *et al.*, 2005), and then we compare actual and expected counts. We wish to detect any *spatial region* (set of nearby locations) where the recent counts for some subset of the data streams monitored are significantly higher than expected: in disease surveillance, this corresponds to an abnormally high incidence of cases of disease in an area, which may indicate an emerging outbreak. We focus here on the univariate case, monitoring a single data stream of counts c^t_i (and the corresponding baselines b^t_i) over multiple spatial locations s_i and time steps t, but our methods generalize to fast multivariate event detection as well. Finally, in the purely spatial case (considering only a single time interval), we omit the superscript t, writing c_i and b_i respectively for the count and baseline of location s_i.

2.1. Methods for spatial event detection

The *spatial and space–time scan statistics* are commonly used methods for event detection (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997). They are in wide use for monitoring health data, detecting clusters of disease cases due to chronic environmental exposures (Kulldorff *et al.*, 1997; Hijalmars *et al.*, 1996), outbreaks of infectious disease (Mostashari *et al.*, 2003) or bioterrorist attacks (Neill, 2006). These methods maximize a score function $F(S)$ over a large set of spatial regions S, each consisting of some subset of locations s_i, and thus can be considered a special case of our general subset scan framework. Typical spatial scan methods constrain the size and shape of the spatial region S and perform an exhaustive search over all regions satisfying the given constraints. Kulldorff's original method (Kulldorff, 1997) assumed circular, purely spatial search regions, but recent variants search for elongated (Neill and Moore, 2004; Kulldorff *et al.*, 2006) or irregular shapes (Duczmal and Assuncao, 2004; Patil and Taillie, 2004; Tango and Takahashi, 2005) and scan over time as an additional search dimension (Kulldorff *et al.*, 1998; Kulldorff, 2001). Finally, the p-value for each region is computed by randomization testing, and any significant regions are reported to the user. Neill *et al.* (2005) developed an *expectation-based* scan statistic which first computes the expected count b^t_i corresponding to each observed count c^t_i by time series analysis, and then compares actual and expected counts. This method adjusts for spatial and temporal variability of the background data, significantly improving the detection time.

Parametric scan statistics (Kulldorff, 1997; Neill *et al.*, 2005; Neill, 2006, 2009) assume some parametric model (such as Poisson- or Gaussian-distributed counts) and maximize the log-likelihood ratio statistic $F(S)$ over all regions S, where

$$F(S) = \log\left[\frac{P\{\text{data}|H_1(S)\}}{P(\text{data}|H_0)}\right].$$

The null hypothesis H_0 assumes no clusters (i.e. all counts are generated from the expected distribution) and the alternative hypothesis $H_1(S)$ assumes that counts in region S are increased by some multiplicative factor. For example, for the expectation-based Poisson (EBP) statistic (Neill *et al.*, 2005), the log-likelihood ratio can be derived as $F(S) = C\ \log(C/B) + B - C$, if $C > B$, and $F(S) = 0$ otherwise, where C and B are respectively the aggregate count Σc^t_i and aggregate baseline Σb^t_i in region S for the given time interval. Similarly, if we assume Gaussian-distributed counts, the log-likelihood ratio for the expectation-based Gaussian (EBG) statistic (Neill, 2006) can be derived as $F(S) = (C' - B')^2/2B'$, if $C' > B'$, and $F(S) = 0$ otherwise, where C' and B' are respectively the aggregate weighted count $\Sigma c^t_i b^t_i/(\sigma^t_i)^2$ and aggregate weighted baseline $\Sigma (b^t_i)^2/(\sigma^t_i)^2$ in region S for the given time interval. As before, c^t_i and b^t_i represent the count

and baseline for location s_i at time step t, and σ_i^t represents the expected standard deviation, also inferred by time series analysis of the historical data for location s_i. Both of these statistics differ from the Poisson spatial scan statistic that was originally proposed by Kulldorff (1997), which compares the ratios of count to baseline inside and outside region S. The log-likelihood ratio for Kulldorff's statistic is defined as

$$F(S) = C \log\left(\frac{C}{B}\right) + (C_{\text{all}} - C) \log\left(\frac{C_{\text{all}} - C}{B_{\text{all}} - B}\right) - C_{\text{all}} \log\left(\frac{C_{\text{all}}}{B_{\text{all}}}\right),$$

if $C/B > C_{\text{all}}/B_{\text{all}}$, and $F(S) = 0$ otherwise, where C and B are defined as above, and C_{all} and B_{all} are the total aggregate count Σc_i^t and baseline Σb_i^t for all spatial locations s_i. Neill (2009) demonstrated that the EBP and EBG statistics have high detection power for both small and large affected regions, whereas Kulldorff's statistic has high detection power for small affected regions but low detection power for large affected regions (Neill, 2009). Other variants of the spatial and space–time scan statistics include the non-parametric (Neill and Lingwall, 2007) and multivariate Bayesian (Neill and Cooper, 2010) scan statistics. These methods have several advantages over typical parametric scan statistics, including the ability to integrate information from multiple sources, to adapt to different data distributions and to distinguish between multiple types of event, but here we focus on the parametric case.

2.2. Accelerating spatial event detection

In the spatial event detection problem, our primary goal is to find the most anomalous spatial or space–time regions (subsets of locations) by efficiently maximizing the score function $F(S)$. Since there are exponentially many subsets to consider ($O(2^N)$ for a spatial data set) with N locations, an exhaustive search over all subsets is typically computationally infeasible. Nevertheless, there are three ways to maximize $F(S)$ efficiently. First, we can reduce the search space, considering only a polynomial number of subsets. For example, Kulldorff's original spatial scan (Kulldorff, 1997) searches over only the $O(N^2)$ distinct circular regions centred at a location; other methods search over rectangles (Neill and Moore, 2004), ellipses (Kulldorff *et al.*, 2006) or cylinders (Kulldorff, 2001). Although such approaches reduce computational complexity, detection power tends to be low for patterns that do not correspond well to the subsets being searched. For example, a search over circles has high power to detect compact clusters but low power to detect elongated or irregular clusters. A second alternative is to search over a larger set of irregular regions, using some heuristic search method to find high scoring subsets. For example, Duczmal and Assuncao (2004) use simulated annealing to search over the space of all connected clusters, whereas Duczmal *et al.* (2007) used a genetic algorithm to maximize a penalized likelihood ratio statistic. The disadvantage of these heuristic search methods is that they are not guaranteed to find a subset which is optimal (maximizes the score function) or even close to optimal.

In this work, we develop new methods which are guaranteed to find the highest scoring subsets of locations *without* an exhaustive search. Neill and Moore (2004) developed a 'fast spatial scan' that can efficiently maximize a score function over the set of rectangular spatial regions, achieving speed-ups of 100–1000 times compared with exhaustive search. However, this method can only be used to maximize Kulldorff's statistic (Kulldorff, 1997) over rectangles, on data aggregated to a uniform grid. The present work enables efficient global optimization of any of a large class of functions, to detect the most interesting subsets of a massive spatial data set. The starting point for this work is our discovery that, for many commonly used spatial scan approaches, including Kulldorff's statistic and many recent variants, we can solve the unconstrained (all subsets) search problem very efficiently. As we discuss below, this LTSS method

also enables us to incorporate proximity constraints, efficiently detecting the most anomalous spatial and space–time regions.

3. Linear time subset scanning

We now formally define the LTSS property and demonstrate that a large class of functions satisfy this property. Let $D = \{R_1 \ldots R_N\}$ be a set of N data records, and let $F(S)$ be a set function mapping a subset of data records $S \subseteq D$ to a real number. For example, in spatial event detection, each region S represents a subset of the spatial locations $\{s_1 \ldots s_N\}$, but our derivations generalize to both spatial and non-spatial data. We refer to F as a 'score function', and $F(S)$ as the 'score' of subset S. Also, let $G(R_i)$ be a function mapping a single data record $R_i \in D$ to a real number. We refer to G as a 'priority function', and $G(R_i)$ as the 'priority' of data record R_i. Next we define $R_{(j)}$, $j = 1, \ldots, N$, to be the data record $R_i \in D$ with the jth highest value of $G(R_i)$. We refer to $R_{(j)}$ as the 'jth highest priority record', and j as the 'priority rank' of record $R_{(j)}$. Given these preliminaries, the LTSS property can be defined as follows.

For a given data set D, the score function $F(S)$ and priority function $G(R_i)$ satisfy the LTSS property if and only if $\max_{S \subseteq D}\{F(S)\} = \max_{j=1\ldots N}[F(\{R_{(1)} \ldots R_{(j)}\})]$.

If the LTSS property holds, we can efficiently maximize $F(S)$ over all subsets of D by evaluating only N of the 2^N possible subsets. If the records $R_1 \ldots R_N$ are already sorted by priority, this property allows us to maximize $F(S)$ in $O(N)$ time, by stepping through the records in priority order and computing the score of each subset $S = \{R_{(1)} \ldots R_{(j)}\}$. Otherwise, we must first sort the records by priority, which requires $O\{N \log(N)\}$ time.

As we demonstrate below, many commonly used score functions, including Kulldorff's original spatial scan statistic and many recently proposed variants, satisfy the LTSS property. We have developed two general proof methods, 'inclusion' and 'substitution', which can be used to show that a given score function $F(S)$ and priority function $G(R_i)$ satisfy LTSS. The simpler case, proof by substitution (which is described in Section 3.4), finds an ordering of records $R_{(1)} \ldots R_{(N)}$ such that substituting a higher priority record $R_{(i)}$ for a lower priority record $R_{(j)}$, where $i < j$, is guaranteed not to decrease the score. In Section 3.1, we prove (by inclusion) that commonly used score functions such as the EBP and EBG scan statistics satisfy the LTSS property with priority function equal to the ratio of count to baseline, $G(s_i) = c_i/b_i$. However, these score functions do not satisfy the substitution property, as shown by the following two examples.

(a) *Example 1*: consider two locations s_1 and s_2, such that s_1 has a higher ratio of count to baseline than s_2, but s_2 has a higher count and baseline, $(c_1, b_1) = (30, 5)$, and $(c_2, b_2) = (100, 50)$. If region S has aggregate count and baseline $(C, B) = (1, 1)$, then $F(S \cup \{s_1\}) > F(S \cup \{s_2\})$ for both EBP and EBG statistics. However, if S has $(C, B) = (100, 100)$, then $F(S \cup \{s_1\}) < F(S \cup \{s_2\})$.

(b) *Example 2*: consider two locations s_1 and s_2, such that s_1 has a higher ratio of count to baseline than s_2, and also has a higher count and baseline, $(c_1, b_1) = (30, 5)$, and $(c_2, b_2) = (2, 1)$. For either of the two regions S that were considered in example 1, $F(S \cup \{s_1\}) > F(S \cup \{s_2\})$ for both EBP and EBG statistics. However, if region S has aggregate count and baseline $(C, B) = (100, 1)$, then $F(S \cup \{s_1\}) < F(S \cup \{s_2\})$.

3.1. Proof of linear time subset scanning property by inclusion

In this section, we describe proof by inclusion and use this method to prove that several commonly used spatial scan statistics satisfy the LTSS property.

(a) For a non-empty subset $S \subset D$, define $R_{\text{in}}(S)$ to be the lowest priority element $R_i \in S$, and define $R_{\text{out}}(S)$ to be the highest priority element $R_i \notin S$. Thus, if $S = \{R_{(1)} \dots R_{(j)}\}$ for some j, then $G\{R_{\text{in}}(S)\} \geqslant G\{R_{\text{out}}(S)\}$, and otherwise $G\{R_{\text{in}}(S)\} \leqslant G\{R_{\text{out}}(S)\}$.

(b) Define $\text{pr}_{\text{in}}(S)$ to be the priority rank of $R_{\text{in}}(S)$, $\text{pr}_{\text{out}}(S)$ to be the priority rank of $R_{\text{out}}(S)$ and $\text{diff}(S) = \text{pr}_{\text{in}}(S) - \text{pr}_{\text{out}}(S)$. We also define $\text{diff}(S) = -1$ for $S = \emptyset$ and $S = D$. Thus, if $S = \{R_{(1)} \dots R_{(j)}\}$ for some j, then $\text{diff}(S) = -1$, and otherwise $\text{diff}(S) > 0$.

(c) Define S^* to be the subset that maximizes $F(S)$, $S^* = \arg\max_S \{F(S)\}$. If there are multiple subsets S^* which maximize $F(S)$, we choose an S^* which *minimizes* $\text{diff}(S^*)$. Thus if $F(S)$ and $G(R_i)$ satisfy the LTSS property then $\text{diff}(S^*) = -1$, and otherwise $\text{diff}(S^*) > 0$.

To prove that a score function $F(S)$ and priority function $G(R_i)$ satisfy the LTSS property we assume that S^* cannot be expressed as $\{R_{(1)} \dots R_{(j)}\}$ for some j. We then show that this leads to a contradiction, by constructing another subset S' with $F(S') \geqslant F(S^*)$ and $\text{diff}(S') < \text{diff}(S^*)$. In particular, we consider the subsets $S_1 = S^* \backslash \{R_{\text{in}}(S^*)\}$, and $S_2 = S^* \cup \{R_{\text{out}}(S^*)\}$. Clearly, $\text{pr}_{\text{in}}(S_1) < \text{pr}_{\text{in}}(S^*)$ and $\text{pr}_{\text{out}}(S_1) = \text{pr}_{\text{out}}(S^*)$, and thus $\text{diff}(S_1) < \text{diff}(S^*)$. Similarly, $\text{pr}_{\text{in}}(S_2) = \text{pr}_{\text{in}}(S^*)$ and $\text{pr}_{\text{out}}(S_2) > \text{pr}_{\text{out}}(S^*)$, and thus $\text{diff}(S_2) < \text{diff}(S^*)$. Since both S_1 and S_2 have lower values of $\text{diff}(S)$ than S^*, a sufficient condition for LTSS is to show that $\max\{F(S_1), F(S_2)\} \geqslant F(S^*)$. In other words, we show that, if the lower priority record R_{in} is included in S^*, then the higher priority record R_{out} must be included as well. As a specific example, we prove that LTSS holds for a large class of score functions which includes the commonly used EBP, EBG and Kulldorff spatial scan statistics.

Theorem 1. Let $F(S) = F(X, Y)$ be a quasi-convex function of two additive sufficient statistics of subset S, $X(S) = \Sigma_{R_i \in S} x_i$ and $Y(S) = \Sigma_{R_i \in S} y_i$, where x_i and y_i depend only on record R_i. Assume that $F(S)$ is monotonically increasing with $X(S)$, and that all y_i-values are positive. Then $F(S)$ satisfies the LTSS property with priority function $G(R_i) = x_i / y_i$.

Proof. We prove theorem 1 by inclusion, defining $S^* = \arg\max_S \{F(S)\}$, $R_{\text{in}}(S)$ and $R_{\text{out}}(S)$ as above. Let $S_1 = S^* \backslash \{R_{\text{in}}(S^*)\}$ and $S_2 = S^* \cup \{R_{\text{out}}(S^*)\}$. We show that $F(S^*) \leqslant \max\{F(S_1), F(S_2)\}$. To do so, define $X^* = \Sigma_{R_i \in S^*} x_i$ and $Y^* = \Sigma_{R_i \in S^*} y_i$. Similarly, define x_{in} and y_{in} to be the x_i- and y_i-values for record $R_{\text{in}}(S^*)$ respectively, and x_{out} and y_{out} to be the x_i- and y_i-values for $R_{\text{out}}(S^*)$ respectively. Thus we must show that $F(X^*, Y^*) \leqslant \max\{F(X^* - x_{\text{in}}, Y^* - y_{\text{in}}), F(X^* + x_{\text{out}}, Y^* + y_{\text{out}})\}$.

The proof proceeds in two steps. First, we show that $F(X^*, Y^*) \leqslant F(X', Y')$, where

$$X' = \frac{y_{\text{out}}}{y_{\text{in}} + y_{\text{out}}}(X^* - x_{\text{in}}) + \frac{y_{\text{in}}}{y_{\text{in}} + y_{\text{out}}}(X^* + x_{\text{out}}) = X^* + \frac{x_{\text{out}} y_{\text{in}} - x_{\text{in}} y_{\text{out}}}{y_{\text{in}} + y_{\text{out}}}$$

and

$$Y' = \frac{y_{\text{out}}}{y_{\text{in}} + y_{\text{out}}}(Y^* - y_{\text{in}}) + \frac{y_{\text{in}}}{y_{\text{in}} + y_{\text{out}}}(Y^* + y_{\text{out}}) = Y^*.$$

Second, we show that $F(X', Y') \leqslant \max\{F(X^* - x_{\text{in}}, Y^* - y_{\text{in}}), F(X^* + x_{\text{out}}, Y^* + y_{\text{out}})\}$.

The first step follows from the assumption that $F(X, Y)$ is monotonically increasing with X, and the facts that $X' \geqslant X^*$ and $Y' = Y^*$. To see that $X' \geqslant X^*$, we note that $G(R_{\text{in}}) \leqslant G(R_{\text{out}})$, and thus $x_{\text{in}}/y_{\text{in}} \leqslant x_{\text{out}}/y_{\text{out}}$. This implies that $x_{\text{out}} y_{\text{in}} - x_{\text{in}} y_{\text{out}}$ is non-negative, and therefore $X' = X^* + (x_{\text{out}} y_{\text{in}} - x_{\text{in}} y_{\text{out}})/(y_{\text{in}} + y_{\text{out}}) \geqslant X^*$. The second step follows from the fact that (X', Y') is a convex combination of $(X^* - x_{\text{in}}, Y^* - y_{\text{in}})$ and $(X^* + x_{\text{out}}, Y^* + y_{\text{out}})$. More precisely, $(X', Y') = \lambda(X^* - x_{\text{in}}, Y^* - y_{\text{in}}) + (1 - \lambda)(X^* + x_{\text{out}}, Y^* + y_{\text{out}})$, where $\lambda = y_{\text{out}}/(y_{\text{in}} + y_{\text{out}})$. The assumption that $F(X, Y)$ is quasi-convex implies that $F(X', Y') \leqslant \max\{F(X^* - x_{\text{in}}, Y^* - y_{\text{in}}), F(X^* + x_{\text{out}}, Y^* + y_{\text{out}})\}$.

Corollary 1. Kulldorff's spatial scan statistic satisfies LTSS, with $G(s_i) = c_i/b_i$. This follows since, for given values of the global count $C_{\text{all}} = \Sigma c_i$ and global baseline $B_{\text{all}} = \Sigma b_i$, we can write $F(S) = F\{C(S), B(S)\}$, where $C(S) = \Sigma_{s_i \in S} c_i$ and $B(S) = \Sigma_{s_i \in S} b_i$. We know that $F(S)$ is monotonically increasing with the count $C(S)$, and all baselines b_i are assumed to be positive. Finally, we prove that $F(C, B)$ is convex (and therefore quasi-convex) by showing that it is the sum of two convex functions, $F_{\text{in}}(C, B) = C \log (C/B)$ and

$$F_{\text{out}}(C, B) = (C_{\text{all}} - C) \log \left(\frac{C_{\text{all}} - C}{B_{\text{all}} - B} \right),$$

and a constant term $-C_{\text{all}} \log(C_{\text{all}}/B_{\text{all}})$. The Hessian of F_{in} is positive semidefinite, with eigenvalues 0 and $1/C + C/B^2 > 0$. The Hessian of F_{out} is also positive semidefinite, with eigenvalues 0 and $1/(C_{\text{all}} - C) + (C_{\text{all}} - C)/(B_{\text{all}} - B)^2 > 0$, and thus F is convex.

3.2. *Linear time subset scanning for separable exponential families*

Although theorem 1 can be used directly to prove that the EBP and EBG scan statistics satisfy the LTSS property, we now prove a stronger result which demonstrates that LTSS holds for many exponential families. Assume that we are given a set of observed counts x_i, the corresponding expected counts μ_i and possibly other parameters, such as the standard deviations σ_i. Let

$$F(S) = \log \left[\frac{P\{\text{data}|H_1(S)\}}{P(\text{data}|H_0)} \right],$$

where the null hypothesis assumes that each observed count x_i is drawn with mean μ_i from a given distribution in a single-parameter exponential family. We can write this distribution in terms of its mean μ as $\log\{P(x|\mu)\} = T(x)\,\theta(\mu) - \psi\{\theta(\mu)\} = T(x)\,\theta(\mu) - \mu\,\theta(\mu) + \phi(\mu)$, where $T(x)$ is the sufficient statistic, $\theta(\mu)$ is a function mapping the mean μ to the natural parameter θ, ψ is the log-partition function and ϕ is the convex conjugate of ψ. For the expectation-based scan statistic (Neill *et al.*, 2005), the alternative hypothesis $H_1(S)$ assumes that counts x_i are drawn with mean $q\mu_i$ inside region S and mean μ_i outside region S, for some constant multiplicative factor $q > 1$. In this case, we can write

$$F(S) = \sup_{q>1} \left(\sum_{s_i \in S} [\log\{P(x_i|q\mu_i)\} - \log\{P(x_i|\mu_i)\}] \right).$$

Plugging in the log-likelihood for the exponential family, we obtain the expression

$$F(S) = \sup_{q>1} \left(\sum_{s_i \in S} [T(x_i)\{\theta(q\mu_i) - \theta(\mu_i)\} + \mu_i\,\theta(\mu_i) - q\mu_i\,\theta(q\mu_i) + \phi(q\mu_i) - \phi(\mu_i)] \right). \quad (1)$$

We now define an exponential family to be *separable* if $\theta(q\mu_i) = z_i\,\theta_0(q) + v_i$, where the function θ_0 depends only on q, whereas z_i and v_i can depend on μ_i and σ_i but are independent of q. We note that the Poisson, Gaussian and exponential distributions are separable (as shown below), but not all exponential families are separable: for example, the binomial and negative binomial distributions $P(x_i|N_i, p_i)$ are only separable with the additional assumption of constant p_i. For a separable exponential family, we can show that $\phi(q\mu_i) = \mu_i z_i\,\phi_0(q) + \mu_i v_i q + K_i$, where $\phi_0(q) = \int \theta_0(q)\,\mathrm{d}q$, and K_i is independent of q. Then the expression for $F(S)$ can be simplified to

$$F(S) = \sup_{q>1} [C\{\theta_0(q) - \theta_0(1)\} + B\{\phi_0(q) - \phi_0(1) + \theta_0(1) - q\,\theta_0(q)\}] \quad (2)$$

where C and B are the sufficient statistics $C = \Sigma_{s_i \in S} T(x_i) z_i$ and $B = \Sigma_{s_i \in S} \mu_i z_i$. To obtain the maximum likelihood estimate of q, we set $\partial F/\partial q = 0$, obtaining $q = C/B$ if $C > B$, and $q = 1$

otherwise. Substituting this value of q into equation (2) and simplifying, we find that $F(S) = B\,D_{\phi_0}(C/B, 1)$ if $C > B$, and $F(S) = 0$ otherwise, where $D_{\phi_0}(x, y)$ is the Bregman divergence, $\phi_0(x) - \phi_0(y) - (x - y)\,\phi_0'(y)$. Thus we can prove the following theorem.

Theorem 2. Let $F(S)$ be the expectation-based scan statistic corresponding to a probability distribution in a separable exponential family. Then $F(S)$ satisfies the LTSS property with priority function $G(s_i) = T(x_i)/\mu_i$.

Proof. As shown above, $F(S) = B\,D_{\phi_0}(C/B, 1)$ if $C > B$, and $F(S) = 0$ otherwise. For $C > B$, the Bregman divergence $D_{\phi_0}(C/B, 1)$ is increasing with C/B, and thus $F(S)$ increases monotonically with C. Additionally, the Hessian of $F(S)$ is positive semidefinite, with eigenvalues 0 and $(1/B + C^2/B^3)\,\theta_0'(C/B) > 0$, and thus $F(S)$ is convex. By theorem 1, $F(S)$ satisfies the LTSS property with priority function $G(s_i) = T(x_i)z_i/\mu_i z_i = T(x_i)/\mu_i$.

Corollary 2. The EBP, EBG and exponential scan statistics all satisfy the LTSS property with priority function $G(s_i) = x_i/\mu_i$. To see this, we note that each distribution belongs to a separable exponential family, with $T(x_i) = x_i$. For each distribution, Table 1 provides the expression for $\theta(q\mu_i)$, its decomposition as $z_i\,\theta_0(q) + v_i$, the sufficient statistics $C = \Sigma_{s_i \in S}\,T(x_i)z_i$ and $B = \Sigma_{s_i \in S}\,\mu_i z_i$, $\phi_0(q)$ and $F(S) = B\,D_{\phi_0}(C/B, 1)$.

Corollary 3. Assume that counts x_i are drawn from a Gaussian distribution with known means μ_i and assumed variances $q_i\sigma_i^2$. We wish to test the null hypothesis H_0: $q_i = 1$ everywhere, against the set of alternative hypotheses $H_1(S)$: $q_i = q$ inside S and $q_i = 1$ outside S, for some constant $q > 1$. This is an expectation-based scan statistic in a separable exponential family, with $T(x_i) = (x_i - \mu_i)^2$ and corresponding expectation $E[(x_i - \mu_i)^2] = \sigma_i^2$. For this distribution, $\theta(q\sigma_i^2) = -1/2q\sigma_i^2 = z_i\,\theta_0(q)$, where $\theta_0(q) = -1/2q$ and $z_i = 1/\sigma_i^2$. Then $\phi_0(q) = -\frac{1}{2}\log(q)$, and $F(S) = B\,D_{\phi_0}(C/B, 1) = \frac{1}{2}\{B\log(B/C) + C - B\}$, where $C = \Sigma_{s_i \in S}\,(x_i - \mu_i)^2/\sigma_i^2$, and $B = \Sigma_{s_i \in S}\,\sigma_i^2/\sigma_i^2 = |S|$. By theorem 2, $F(S)$ satisfies the LTSS property with $G(s_i) = (x_i - \mu_i)^2/\sigma_i^2 = Z_i^2$, the squared z-score corresponding to observed count x_i.

We note that theorem 2 assumes the standard, one-sided version of the expectation-based scan statistic, where $F(S)$ is positive if $C > B$ and 0 otherwise. This statistic is commonly used to detect spatial clusters of increased counts (e.g. emerging clusters of disease cases). To detect decreased counts, we can define $F(S)$ to be positive if $C < B$ and 0 otherwise. In this case, we write the statistic as $F(B, C)$ and note that $F(S)$ is convex and monotonically increasing with B, and thus the statistic satisfies the LTSS property with priority function $G(s_i) = \mu_i/T(x_i)$. Finally, the two-sided statistic (detecting either increased or decreased counts) can be efficiently optimized by maximizing both one-sided statistics, and then taking the maximum of the two results.

The observation that exponential family score functions $F(S)$ can be written in terms of a Bregman divergence, and can be proven to satisfy the LTSS property, suggests the question

Table 1. Derivation of $F(S)$ for expectation-based scan statistics in a separable exponential family

Distribution	$\theta(q\mu_i)$	$\theta_0(q)$	z_i	v_i	C	B	$\phi_0(q)$	$F(S)$
Poisson	$\log(q\mu_i)$	$\log(q)$	1	$\log(\mu_i)$	$\sum_{s_i \in S} x_i$	$\sum_{s_i \in S} \mu_i$	$q\log(q) - q$	$C\log\left(\frac{C}{B}\right) + B - C$
Gaussian	$\frac{q\mu_i}{\sigma_i^2}$	q	$\frac{\mu_i}{\sigma_i^2}$	0	$\sum_{s_i \in S} \frac{x_i\mu_i}{\sigma_i^2}$	$\sum_{s_i \in S} \frac{\mu_i^2}{\sigma_i^2}$	$\frac{q^2}{2}$	$\frac{(C-B)^2}{2B}$
Exponential	$-\frac{1}{q\mu_i}$	$-\frac{1}{q}$	$\frac{1}{\mu_i}$	0	$\sum_{s_i \in S} \frac{x_i}{\mu_i}$	$\sum_{s_i \in S} \frac{\mu_i}{\mu_i} = \lvert S\rvert$	$-\log(q)$	$B\log\left(\frac{B}{C}\right) + C - B$

of whether *all* Bregman divergences satisfy the LTSS property. The answer to this question is no: whereas the Bregman divergences $D_\phi(C/B, 1)$ are convex, general Bregman divergences $D_\phi(C, B)$ are convex in their first argument but not necessarily in their second argument, and non-convex Bregman divergences are not guaranteed to satisfy the LTSS property. As a counterexample, we consider the Bregman divergence

$$F(S) = D_\phi(C, B) = \frac{(C-B)^2}{(C+1)(B+1)^2},$$

corresponding to the convex function $\phi(x) = 1/(x+1)$. Given the three locations s_1, s_2 and s_3, where $(c_1, b_1) = (1.95, 1.35)$, $(c_2, b_2) = (4.25, 2.90)$ and $(c_3, b_3) = (1.00, 0.65)$, assume that there is a priority function $G(s_i)$ such that $F(S)$ satisfies the LTSS property with priority function $G(s_i)$. For the given locations and score function, the highest scoring subset of $\{s_1, s_2\}$ is $\{s_2\}$, and thus we must have $G(s_2) > G(s_1)$. However, the highest scoring subset of $\{s_1, s_2, s_3\}$ is $\{s_1, s_3\}$, and thus we must have $G(s_1) > G(s_2)$. This is a contradiction, and thus no priority function $G(s_i)$ exists such that $F(S)$ satisfies the LTSS property.

3.3. Strong linear time subset scanning

Some score functions $F(S)$ and associated priority functions $G(R_i)$ allow us to prove a stronger property, which enables efficient maximization of $F(S)$ over all subsets $S \subseteq D$ with a given cardinality j. This property, 'strong LTSS', is defined as follows.

For a given data set D, the score function $F(S)$ and priority function $G(R_i)$ satisfy the strong LTSS property if and only if, for all $j \in \{1 \ldots N\}$, $\max_{S \subseteq D:|S|=j}\{F(S)\} = F(\{R_{(1)} \ldots R_{(j)}\})$.

The LTSS property that was defined in the previous section can be called 'weak LTSS', to distinguish it from the strong LTSS property that is defined here. Clearly, strong LTSS implies weak LTSS, since $\max_{S \subseteq D}\{F(S)\} = \max_{j=1\ldots N} \max_{S \subseteq D:|S|=j}\{F(S)\}$, and strong LTSS allows efficient maximization for each value of j. Thus any score function $F(S)$ satisfying strong LTSS can also be efficiently maximized over all subsets of the data.

However, we note that weak LTSS does not imply strong LTSS, and that in fact the EBP, EBG and Kulldorff spatial scan statistics (which were shown above to satisfy the weak LTSS property) *do not* satisfy the strong LTSS property. As a simple counterexample, we consider the EBP statistic, with $F(S) = C \log(C/B) + B - C$ and $G(s_i) = c_i/b_i$ defined as above. Given a data set of two spatial locations where $s_{(1)}$ has $(c_i, b_i) = (10, 1)$ and $s_{(2)}$ has $(c_i, b_i) = (100, 50)$, the highest scoring 1-element subset is $\{s_{(2)}\}$, not $\{s_{(1)}\}$, so strong LTSS does not hold. Similar counterexamples can be constructed for the EBG and Kulldorff statistics as well.

3.4. Proof of strong linear time subset scanning property by substitution

In this section, we describe proof by substitution and use this method to prove that several commonly used spatial scan statistics satisfy the strong LTSS property. As above, for a non-empty subset $S \subset D$, we define $R_{\text{in}}(S)$ to be the lowest priority element contained in S, $R_{\text{out}}(S)$ to be the highest priority element *not* contained in S, $\text{pr}_{\text{in}}(S)$ to be the priority rank of $R_{\text{in}}(S)$, $\text{pr}_{\text{out}}(S)$ to be the priority rank of $R_{\text{out}}(S)$ and $\text{diff}(S) = \text{pr}_{\text{in}}(S) - \text{pr}_{\text{out}}(S)$. Additionally, we define S_j^* to be the subset that maximizes $F(S)$ among all subsets of cardinality j: $S_j^* = \arg\max_{S:|S|=j}\{F(S)\}$. If there are multiple subsets S_j^* which maximize $F(S)$, we choose an S_j^* which *minimizes* $\text{diff}(S_j^*)$. Thus, if $F(S)$ and $G(R_i)$ satisfy the strong LTSS property, then $\text{diff}(S_j^*) = -1$ for all j, and otherwise $\text{diff}(S_j^*) > 0$ for some j.

To prove that a score function $F(S)$ and priority function $G(R_i)$ satisfy the strong LTSS property, we assume that, for some $j \in \{1 \ldots N\}$, S_j^* is not equal to $\{R_{(1)} \ldots R_{(j)}\}$. We then show

that this leads to a contradiction, by constructing another subset S'_j with $|S'_j| = j$, $F(S'_j) \geq F(S^*_j)$ and $\text{diff}(S'_j) < \text{diff}(S^*_j)$. In particular, we consider the subset $S'_j = S^*_j \cup \{R_{\text{out}}(S^*_j)\} \setminus \{R_{\text{in}}(S^*_j)\}$. Clearly, $\text{pr}_{\text{in}}(S'_j) < \text{pr}_{\text{in}}(S^*_j)$ and $\text{pr}_{\text{out}}(S'_j) > \text{pr}_{\text{out}}(S^*_j)$, and thus $\text{diff}(S'_j) < \text{diff}(S^*_j)$. Since S'_j has a lower value of $\text{diff}(S)$ than S^*_j, a sufficient condition for strong LTSS is to show that $F(S'_j) \geq F(S^*_j)$. In other words, we show that the score of subset S^*_j would not be reduced if we *substitute* R_{out} for R_{in}. We prove that the strong LTSS property holds for a large class of functions.

Theorem 3. Let $F(S) = F(X, |S|)$ be a function of one additive sufficient statistic of subset S, $X(S) = \Sigma_{R_i \in S} x_i$ (where x_i depends only on record R_i), and the cardinality of S. Assume that $F(S)$ is monotonically increasing with X. Then $F(S)$ satisfies the strong LTSS property with priority function $G(R_i) = x_i$.

Proof. We prove theorem 3 by substitution, defining S^*_j, $R_{\text{in}}(S)$ and $R_{\text{out}}(S)$ as above. Assume that, for some $j \in \{1 \ldots N\}$, $S^*_j \neq \{R_{(1)} \ldots R_{(j)}\}$. Let $S'_j = S^*_j \cup \{R_{\text{out}}(S^*_j)\} \setminus \{R_{\text{in}}(S^*_j)\}$ as above. We show that $F(S'_j) \geq F(S^*_j)$. To do so, define $X^* = \Sigma_{R_i \in S^*_j} x_i$, and define x_{in} and x_{out} to be the x_i-values for records $R_{\text{in}}(S^*_j)$ and $R_{\text{out}}(S^*_j)$ respectively. Thus we must show that $F(X^* - x_{\text{in}} + x_{\text{out}}) \geq F(X^*)$. This follows from the assumption that $F(X)$ is monotonically increasing with X, and the fact that $G(R_{\text{out}}) = x_{\text{out}}$ is greater than or equal to $G(R_{\text{in}}) = x_{\text{in}}$, and therefore $X^* - x_{\text{in}} + x_{\text{out}} \geq X^*$.

Corollary 4. The expectation-based scan statistics for the exponential distribution (corollary 2) and variance of a Gaussian distribution (corollary 3) satisfy the strong LTSS property. In each case, we can write $F(S)$ is proportional to $|S| \log(|S|/C) + C - |S|$ if $C > |S|$ and 0 otherwise, where C is an additive sufficient statistic of subset S. For $C > |S|$, we observe that $\partial F / \partial C = 1 - |S|/C > 0$. Thus $F(S)$ is monotonically increasing with C and satisfies the strong LTSS property.

For score functions $F(S)$ satisfying the strong LTSS property by theorem 3, we can include a penalty term that is a function of $|S|$, and the resulting penalized score function $F(S) - H(|S|)$ also satisfies the strong LTSS property. This is useful because multiplicity considerations suggest that the unpenalized subset scan is biased towards detecting subsets with $|S| \approx N/2$, since $N!/|S|!(N - |S|)!$ subsets of cardinality $|S|$ are considered. However, many score functions $F(S) = F(C, B)$ do not satisfy the strong LTSS property unless the baselines b_i are constant, in which case $B = \Sigma_{s_i \in S} b_i \propto |S|$. If $F(S)$ satisfies only the weak LTSS property, then $F(S) - H(|S|)$ is not guaranteed to satisfy the LTSS property. Below, we propose an alternative approach (penalizing the neighbourhood size) which implicitly penalizes the cardinality of S as well as penalizing spatially dispersed clusters.

3.5. Extensions to space–time and multivariate data

Since the parametric scan statistics are functions of the additive sufficient statistics of region S, we can easily extend our proofs of the LTSS property to the space–time case (Kulldorff, 2001), where we scan over varying temporal windows consisting of the most recent W time steps, for $W = 1 \ldots W_{\max}$. Let $F(S) = \max_{W=1 \ldots W_{\max}} \{F_W(S)\}$, where $F_W(S)$ assumes a fixed temporal window size W. Then $F_W(S) = F(C_W, B_W)$, where $C_W(S) = \Sigma_{s_i \in S} \Sigma_{t=0 \ldots W-1} c_i^t = \Sigma_{s_i \in S} c_{i,W}$ and $B_W(S) = \Sigma_{s_i \in S} \Sigma_{t=0 \ldots W-1} b_i^t = \Sigma_{s_i \in S} b_{i,W}$. We can efficiently maximize $F_W(S)$ for each temporal window size W, by computing the aggregate count $c_{i,W}$ and baseline $b_{i,W}$ for each location, and then prioritizing the locations by $G(s_i) = c_{i,W}/b_{i,W}$.

Extension of the LTSS property to the multivariate case is likewise straightforward in cases when the multivariate statistic is a function of one or more additive sufficient statistics, aggre-

gated over all locations, time steps and data streams. In this case, we can efficiently optimize the score function $F(S)$ for a fixed set of data streams $D \subseteq \{D_1 \ldots D_M\}$ and a fixed temporal window W. For multiple subsets of streams, or varying temporal window sizes, we can perform a separate optimization for each, and then maximize over all subsets of streams and temporal window sizes under consideration. Here we consider the original multivariate formulation of the spatial scan statistic that was proposed by Burkom (2003), which aggregates counts and baselines across the multiple data streams being monitored and applies the univariate scan statistic to these aggregates. This approach is distinct from Kulldorff's multivariate spatial scan statistic (Kulldorff *et al.*, 2007), which assumes that data streams are conditionally independent and thus adds the log-likelihood ratio scores across the multiple streams. For Burkom's method, the parametric scan statistic is a function of the aggregate count and aggregate baseline of region S for the given data streams and temporal window, and thus the LTSS property holds with $G(s_i) = c_i/b_i$, where $c_i = \Sigma_{D_m \in D} \Sigma_{t=0 \ldots W-1} c^t_{i,m}$ and $b_i = \Sigma_{D_m \in D} \Sigma_{t=0 \ldots W-1} b^t_{i,m}$.

Whereas LTSS allows efficient optimization over subsets of locations for each subset of streams, optimization of the scan statistic over all subsets of M streams requires time proportional to 2^M and thus is computationally infeasible when the number of streams is large. However, an alternative approach is to consider a relatively small number of spatial regions (e.g. searching over circles rather than all subsets of locations). For any given spatial region S (and a fixed temporal window W), we can efficiently optimize over all subsets of streams. To do so, we can order the M data streams D_m by a priority function $G(D_m)$ and consider only the k highest priority streams for each $k = 1, \ldots, M$. Then we have $G(D_m) = c_m/b_m$, where $c_m = \Sigma_{s_i \in S} \Sigma_{t=0 \ldots W-1} c^t_{i,m}$ and $b_m = \Sigma_{s_i \in S} \Sigma_{t=0 \ldots W-1} b^t_{i,m}$. Thus we can use the LTSS property either to optimize efficiently over subsets of locations for a given subset of streams, or to optimize efficiently over subsets of streams for a given subset of locations. Further investigation of the multivariate case is beyond the scope of this paper, and we focus on the univariate case for the remainder of our discussion.

3.6. Comparison with related methods

As noted in Section 1, our unconstrained fast subset scan approach, based on the LTSS property, is similar to the ULS scan statistic that was proposed by Patil and Taillie (2004), which has been widely applied to graph and network data. The ULS approach also orders the spatial locations by priority, where the priority function is defined as $G(s_i) = c_i/b_i$, and considers the top k highest priority locations for each $k = 1, \ldots, N$. Rather than considering the subset consisting of all k locations, however, the ULS enforces a connectivity constraint, considering the connected components of the subgraph formed by the top k locations for each k. Thus, for a fully connected graph, the ULS reduces to the unconstrained fast subset scan approach.

The contributions of this paper, compared with the original ULS approach, are twofold. First, whereas Patil and Taillie (2004) focused on the specific case of optimizing Kulldorff's univariate spatial scan statistic with connectivity constraints, our fast subset scan approach can optimize a large class of score functions for multivariate spatial, space–time and non-spatial data. Moreover, the unconstrained fast subset scan can be used as a building block to solve a wide variety of constrained subset scan problems, e.g. incorporating hard or soft constraints on spatial proximity. Second, we *prove* that the fast subset scan is guaranteed to optimize efficiently and exactly any score function which satisfies the LTSS property. Patil and Taillie neither proved nor claimed that the ULS is guaranteed to maximize Kulldorff's statistic over the set of connected regions and, in fact, a simple counterexample demonstrates that the ULS may find a suboptimal region. Consider a four-node 'Y-junction' graph, with nodes s_1, s_2 and s_3 each connected to the centre node s_4. Let $(c_1, b_1) = (c_2, b_2) = (10, 1)$, $(c_3, b_3) = (10, 10)$ and $(c_4, b_4) = (0, 1)$. Then

the ULS would consider only the subsets $\{s_1\}$, $\{s_2\}$, $\{s_3\}$ and $\{s_1, s_2, s_3, s_4\}$, failing to identify the highest scoring connected subset $\{s_1, s_2, s_4\}$. We have recently developed the GraphScan method (Speakman and Neill, 2010), which incorporates connectivity constraints into the LTSS framework and is guaranteed to find the highest scoring connected subset. However, a detailed discussion of GraphScan is beyond the scope of this paper.

We also note that the LTSS property is distinct from prior work in submodular function optimization. Submodular functions have an intuitive 'diminishing marginal returns' property and can be approximately maximized by greedy search (Nemhauser *et al.*, 1978), enabling near optimal solutions to problems such as feature selection and sensor placement (Leskovec *et al.*, 2007). However, LTSS enables us to find efficiently an exact, rather than approximate, solution. Moreover, a simple example demonstrates that LTSS can be applied to functions that are neither submodular nor supermodular. Consider the EBP statistic, given three spatial locations s_1, s_2 and s_3 where $(c_1, b_1) = (3, 1)$ and $(c_2, b_2) = (c_3, b_3) = (2, 1)$. Then $F(\{s_1, s_3\}) - F(\{s_1\}) < F(\{s_1, s_2, s_3\}) - F(\{s_1, s_2\})$, so F is not submodular. However, $F(\{s_2, s_3\}) - F(\{s_2\}) > F(\{s_1, s_2, s_3\}) - F(\{s_1, s_2\})$, so F is not supermodular. We can also consider the variant of EBP which assumes uniform baselines, $F(S)$ equals $C \log(C/|S|) + |S| - C$ if $C > |S|$, and 0 otherwise. This function satisfies the strong LTSS property, but identical calculations demonstrate that it is neither submodular nor supermodular.

3.7. *Initial evaluation for spatial and space–time data*

As a concrete example of the utility of LTSS in practice, we considered a spatial disease surveillance data set consisting of the daily counts of emergency department visits with respiratory symptoms (cough and shortness of breath) in 97 Allegheny County zip codes. This data set is described in detail in Section 5.1 below. An exhaustive search over the 2^{97} possible subsets of zip codes would be computationally infeasible, requiring over 10^{20} years of computation time for a single day of data. However, LTSS enabled efficient maximization of $F(S)$ for the EBP, EBG and exponential scan statistics, as well as Kulldorff's spatial scan statistic, requiring approximately 0.04 s per day of data for spatial and space–time scans with temporal window sizes up to $W = 28$.

Although these results demonstrate the potential of LTSS to enable efficient unconstrained maximization of the score function for real world spatial and space–time data, we note that unconstrained maximization over subsets is typically not sufficient to solve practical spatial detection problems. Since our search over subsets does not take the spatial proximity of locations into consideration, the highest scoring 'region' may consist of a dispersed set of locations, e.g. one zip code in the north-west corner and one zip code in the south-east corner of the county. In the following section, we consider how spatial constraints can be incorporated into LTSS to enable efficient maximization over regions that are constrained by spatial proximity.

4. Incorporating spatial constraints

To incorporate spatial information in the subset scan framework, assume that we are given a metric which specifies the distance $d(s_i, s_j)$ between any two spatial locations s_i and s_j. We then maximize $F(S)$ over only those subsets which satisfy some constraint on *proximity*, e.g. an upper bound on the maximum distance between locations. Here we propose an efficient proximity-constrained subset scan method which we call 'fast localized scan'. The fast localized scan approach considers each spatial location s_i, $i = 1, \ldots, N$, as a possible 'centre' of the region.

For each centre location s_i, we consider its 'local neighbourhood' S_i and use LTSS to maximize efficiently over all subsets $S \subseteq S_i$.

(a) For the 'fixed neighbourhood' (fixed k) approach to the fast localized scan, we define the local neighbourhood S_i to consist of the centre location s_i and its $k-1$ nearest neighbours.
(b) For the 'fixed radius' (fixed r) approach, we define S_i to consist of the centre location s_i and all other locations within distance r of the centre.

In either case, assuming that the local neighbourhood of a given centre s_i contains k locations, LTSS allows us to maximize $F(S)$ for the given neighbourhood by evaluating only $O(k)$ of the $O(2^k)$ subsets. Assuming that the k locations have already been sorted by priority, we need only to evaluate the subsets consisting of the j highest priority locations, for $j = 1, \ldots, k$. This results in a total run time of $O\{Nk + N\log(N)\}$ for the fixed k approach, and $O\{N\bar{k} + N\log(N)\}$ for the fixed r approach, where $\bar{k}$ is the average neighbourhood size corresponding to the fixed radius r. In these expressions, the additional $O\{N\log(N)\}$ term results from sorting the N locations by priority, which needs only to be done once (rather than once per centre). This analysis also assumes that the k nearest neighbours have been precomputed for each location.

We note that the fixed k fast localized scan is very similar to the flexible spatial scan statistic (FlexScan) that was proposed by Tango and Takahashi (2005), in that it searches over subsets of neighbourhoods defined by a centre location and its $k-1$ nearest neighbours. The two fundamental differences are that FlexScan requires the resulting region to be connected, whereas the fast localized scan can return a disconnected region if it satisfies the proximity constraints, and that the run time of FlexScan scales exponentially rather than linearly with k, making it computationally infeasible for $k > 30$ (Tango and Takahashi, 2005).

In Fig. 1, we show the total computation time required to optimize the EBP statistic over proximity-constrained subsets (as a function of the neighbourhood size k) for 100 days of data. We compare the LTSS-enabled fast localized scan with a 'naive localized scan' which does not use LTSS; the run time of FlexScan is also shown for comparison. With LTSS, the run time increased linearly with neighbourhood size, up to a maximum of 5.0 s. Without LTSS, the run time increased exponentially with neighbourhood size, requiring approximately 50 h for 100 days of data at $k = 25$ and nearly 2 years for a single day of data at $k = 40$. The run time of FlexScan also increased exponentially with neighbourhood size, since it performs a separate connectivity check for each of the $O(2^k)$ regions centred at each location (Tango and Takahashi, 2005). Similarly, in the fixed radius case, performing a localized scan without LTSS

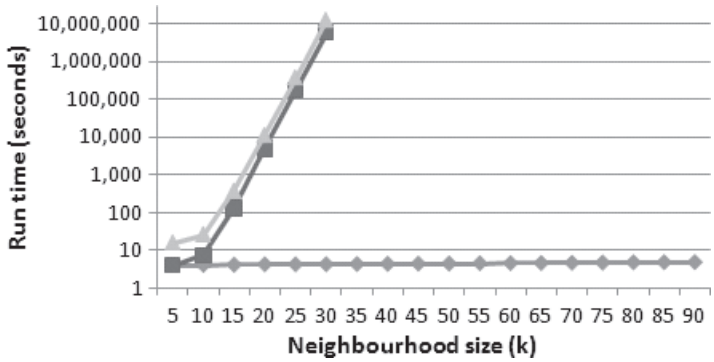

Fig. 1. Performance results for the fixed k fast localized scan, as a function of the neighbourhood size k: total run times for 100 days of data, for the EBP statistic with (◆) and without (■) LTSS; run times for Tango and Takahashi's (2005) flexible scan statistic FlexScan (▲) are also shown for comparison

required approximately 6 min of run time for 100 days of data at $r = 0.10$ (with distances measured in degrees latitude and degrees longitude) and was computationally infeasible for $r = 0.20$, whereas LTSS enabled computation for any value of r in under 5 s for 100 days of data. These experiments were repeated using the EBG and Kulldorff scan statistics, and the run times were nearly identical.

4.1. Fast localized multiscan

Although the fixed neighbourhood and fixed radius scans enable efficient maximization of the score function $F(S)$ subject to proximity constraints, these methods have two distinct disadvantages. First, because they enforce a hard constraint on the maximum size of region (neighbourhood k or radius r), they are likely to lose detection power and spatial accuracy whenever the size of the affected region is larger than this constraint. Second, because all subsets satisfying the proximity constraint are considered 'equally likely' (i.e. larger subsets are not penalized), a large value of the maximum neighbourhood size k or radius r can cause the method to report spatially dispersed subsets that we would not typically consider to be a single spatial region. An alternative approach is to choose a region which maximizes some function of the score $F(S)$ and the size of region S. Unfortunately, as noted above, most arbitrarily chosen functions of score and size may not satisfy the LTSS property. However, we can separately compute the highest scoring subset for each neighbourhood size $k = 1, \ldots, N$, and then choose the subset which optimizes the trade-off between score and size. Since LTSS allows us to compute the highest scoring subset very quickly for each value of k, it makes this 'fast localized multiscan' approach computationally feasible.

More precisely, the fast localized multiscan consists of the following steps. For each centre location s_i and each neighbourhood size $k = 1, \ldots, N$, we define S_{ik} to be the set of locations consisting of s_i and its $k-1$ nearest neighbours, and we use LTSS to maximize $F(S)$ efficiently over all subsets of S_{ik} in $O(k)$ time. For each S_{ik}, we record the highest scoring subset $S^* = \arg\max_{S \subseteq S_{ik}} \{F(S)\}$, its score $F^* = F(S^*)$, the neighbourhood size k and the radius r (the distance from s_i to its $(k-1)$th nearest neighbour). Given this set of $O(N^2)$ regions, we then form the *Pareto set* consisting of all regions which optimize the trade-off between score and neighbourhood size. We exclude any region S which is *dominated* by another smaller and higher scoring region.

(a) For the 'multiscan k' approach, the Pareto set consists of all subsets S such that no other subset S' has either $F(S') > F(S)$ and $k(S') \leqslant k(S)$, or $F(S') = F(S)$ and $k(S') < k(S)$.
(b) For the 'multiscan r' approach, the Pareto set consists of all subsets S such that no other subset S' has either $F(S') > F(S)$ and $r(S') \leqslant r(S)$, or $F(S') = F(S)$ and $r(S') < r(S)$.

Finally, we can choose a single region S from the Pareto set, on the basis of the desired trade-off between score and neighbourhood size. For example, for the multiscan k approach, we can choose the region that maximizes $F(S) - Lk$ and, for the multiscan r approach, we can choose the region that maximizes $F(S) - Lr$, for some constant L. The run time of the fast multiscan is $O(N^3)$, since we must evaluate N centres and N neighbourhood sizes $k = 1, \ldots, N$, and each optimization can be performed in $O(k)$ time. If we restrict the maximum neighbourhood size to some constant $k_{\max}$, then the run time is reduced to $O\{Nk_{\max}^2 + N \log(N)\}$, where the $O\{N \log(N)\}$ term results from the initial sorting of locations by priority.

Performance results for the fast multiscan, as a function of the maximum neighbourhood size $k_{\max}$, are shown in Fig. 2. For the EBP statistic, the fast multiscan required a total run time up to 76.5 s for 100 days of emergency department data. Similar results were seen for the EBG statistic (87.7 s) and Kulldorff's statistic (88.8 s), demonstrating that the fast multiscan can

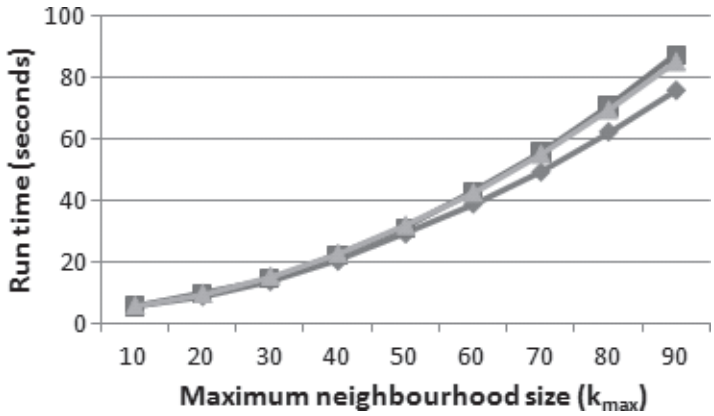

Fig. 2. Performance results for the fast multiscan, as a function of the maximum neighbourhood size $k_{\max}$: total run time for 100 days of data for the EBP (◆), EBG (▲) and Kulldorff (■) statistics with LTSS

find the spatial region which optimizes the trade-off between likelihood ratio score and spatial proximity in less than 1 s per day of data.

5. Evaluation

We now present an empirical comparison of detection time and spatial accuracy for our fast localized scan and multiscan methods, using a large set of simulated respiratory disease outbreaks injected into real world emergency department data from Allegheny County, Pennsylvania. We compared six variants of spatial scan: 'circles' (the traditional Kulldorff approach, searching over the set of circular regions centred at each spatial location), 'all subsets' (using LTSS without proximity constraints), fixed k and fixed r fast localized scans, and multiscan k and multiscan r fast localized multiscans. For each method (except for circles and all subsets), we considered 12 distinct parameter settings. For the fixed k method, we used neighbourhood sizes of $k = 5, 10, \ldots, 60$. For the fixed r method, we used radii of $r = 0.02, 0.04, \ldots, 0.24$ (all spatial co-ordinates were given in degrees latitude and degrees longitude). For the multiscan k method, we used weights $L = 0.1, 0.2, \ldots, 1.2$ to choose a region from the Pareto set and, for the multiscan r method, we used weights $L = 20, 40, \ldots, 240$. The range of parameters for each method was chosen to cover the entire continuum from very strong proximity constraints (where large neighbourhoods are disallowed or severely penalized) to very weak proximity constraints (approximating an unconstrained subset scan). For each method, we used the EBP space–time scan statistic, with a maximum temporal window size of $W_{\max} = 3$. We now describe the data, outbreak simulations, evaluation metrics and results in detail.

5.1. Description of emergency department data

We obtained a data set of 612713 deidentified emergency department visit records collected from 10 Allegheny County hospitals from January 1st, 2004, to December 31st, 2005. Each record contains fields for the patient's date of admission to the emergency department, home zip code, chief complaint (free text) and international classification of diseases, version 9, 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 397134 records (64.8%). The free-text chief complaint was present for all remaining records, and the international classifi-

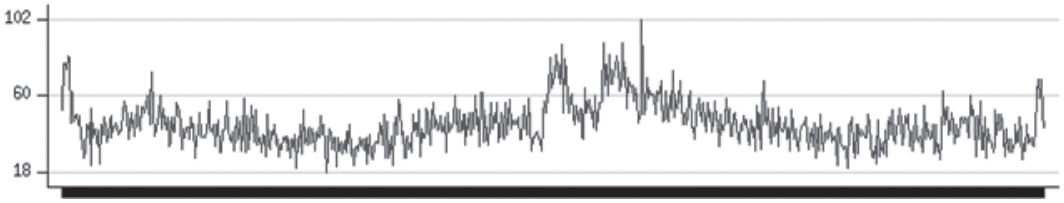

Fig. 3. Daily counts of Allegheny County emergency department cases with respiratory symptoms from January 1st, 2004, to December 31st, 2005

cation of diseases code was present for 336338 (84.7%) of the remaining records. From these data, we created a count data set by recording the number of patient records with respiratory symptoms in each zip code for each day. A patient record was determined to exhibit respiratory symptoms if its chief complaint string contained the substrings 'cough', 'dyspnoea', 'shortness' or 'sob', or if its international classification of diseases code was equal to 786.2 (cough) or 786.05 (shortness of breath). The set of records was then manually refined to remove spurious substring matches. The resulting respiratory emergency department count data set had a mean daily count of 44.0 cases, with a standard deviation of 12.1 cases. The time series of daily counts (aggregated over all Allegheny County zip codes) is shown in Fig. 3. Since cases were spread over 97 zip codes, many zip codes had zero counts on any given day. The data exhibited slight but statistically significant day-of-week trends, with counts peaking on Mondays, and clear seasonal trends, with counts peaking in February.

We note that the use of real rather than simulated background data has several advantages. The real world data incorporate seasonal and day-of-week trends, as well as spatial heterogeneity, which would not be present in typical simulated data sets. A successful disease surveillance system should be able to detect outbreaks reliably without producing an excessive number of false positive alarms due to the variability in the background data, and thus we believe that our semisynthetic simulation approach will produce more relevant evaluation results than typical fully synthetic simulations. However, one drawback to our approach is the possible presence of true disease outbreaks in the background data, which could skew our evaluation results. Thus, as a check of robustness, we performed all simulations twice: once by using the original count data, and once by using simulated counts redrawn from a Poisson distribution which preserved the spatial and temporal trends in the data, but assumed independence of space and time. More precisely, each simulated count c_i^t was redrawn from a Poisson distribution with mean

$$\mu_i^t = \sum_i c_i^t \sum_t c_i^t \Big/ \sum_i \sum_t c_i^t,$$

where the sums were taken over all locations s_i and all time steps t. Overall performance results were very similar for the two background count distributions, suggesting that the presence of true outbreaks or other sources of space–time interaction in the real world data did not substantially affect the results. Thus we focus on the real world data set for the remainder of our discussion.

5.2. Simulation of outbreaks

We used a semisynthetic testing framework (injecting simulated respiratory outbreaks into the real world emergency department data) to compare the detection power and spatial accuracy of our 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 duration of outbreak T, the severity of outbreak Δ

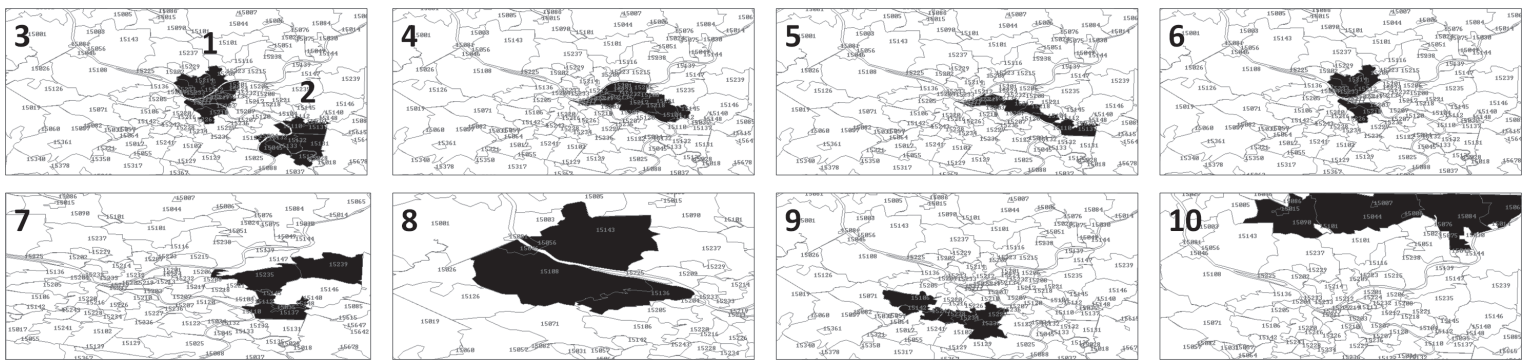

Fig. 4. 10 simulated outbreak regions used in our semisynthetic tests: outbreak region 3 consists of two disjoint circular clusters; outbreak region 1 is the north-west cluster only and outbreak region 2 is the south-east cluster only

and the subset of affected zip codes S_{inject}. 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($tw_i\Delta$) cases into each affected zip code, where w_i is the 'weight' of that zip code,

$$w_i = \sum_t c_i^t \Big/ \sum_i \sum_t c_i^t.$$

We considered 10 differently shaped outbreak regions S_{inject}, including approximately equal numbers of circular, elongated and irregular regions, as shown in Fig. 4. All outbreaks were assumed to be 2 weeks in duration ($T = 14$), and we assumed $\Delta = 1$. For each region, we created 200 different, randomly generated outbreaks, giving a total of 2000 outbreaks for evaluation.

We note that simulation of outbreaks is an active area of on-going research in biosurveillance. The creation of realistic outbreak scenarios is important because of the difficulty of obtaining sufficient labelled data from real outbreaks, but it is also very challenging. State of the art outbreak simulations such as those of Buckeridge *et al.* (2004), Wallstrom *et al.* (2005) and Hogan *et al.* (2007) 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 the duration of outbreak and severity. Although 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.

5.3. Comparison of detection power

For each method, 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 data set with no outbreaks injected. Then, for each of the 2000 injected outbreaks, we computed the maximum region score for each outbreak day. For a given false positive rate r, the 'days to detect' for a given outbreak are computed as the first day of outbreak ($t = 1, \ldots, 14$) with maximum region score higher than the $100(1-r)$ percentile of the maximum region scores for the original data set. If no day of the outbreak has score higher than this threshold, 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.

Fig. 5 shows the average time to detect for each method, at a fixed false positive rate of 1 per month. Results were averaged over all 2000 outbreaks. Searching over circular regions, as in the original spatial scan approach (Kulldorff, 1997), required an average of 9.43 days to detect, with 79.3% of outbreaks detected. Searching over all subsets (without spatial proximity constraints) required an average of 10.25 days to detect, with 70.8% of outbreaks detected. The fixed k fast localized scan achieved faster detection than the circles approach for all values of k between 5 and 45, requiring a minimum of 7.60 days to detect for $k = 10$. The fixed r fast localized scan achieved faster detection than the circles approach for all values of r between 0.02 and 0.14, requiring a minimum of 7.64 days to detect for $r = 0.06$. The fast multiscan k method achieved faster detection than the circles approach for all values of the L-parameter between 0.2 and 1.2, requiring a minimum of 7.54 days to detect for $L = 0.8$. Finally, the fast multiscan r method achieved faster detection than the circles approach for all values of the L-parameter between 120 and 240, requiring a minimum of 7.59 days to detect for $L = 220$. All four methods could detect over 90% of outbreaks for the given parameters. These results demonstrate improved performance of all four LTSS-based spatial scan methods compared with Kulldorff's circular scan, across a

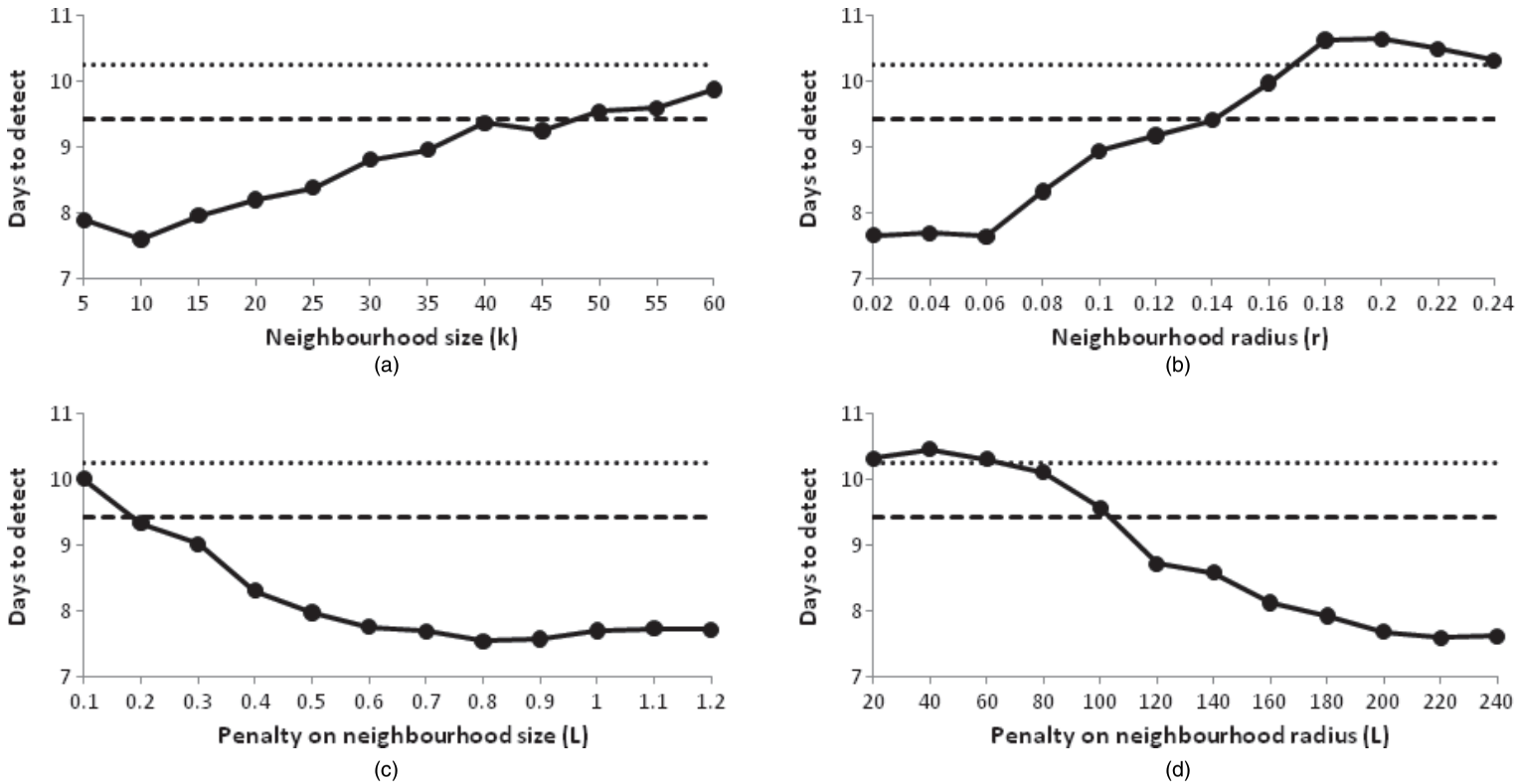

Fig. 5. Comparison of detection methods: average number of days to detect at one false positive alarm per month (———), compared with the circles (– – –) and all-subsets (·····) approaches for (a) the fixed k, (b) fixed r, (c) multiscan k and (d) multiscan r methods

wide range of parameter values. For well-chosen parameters, each of these methods could detect nearly 2 days faster than searching over circles, with fewer than half as many missed outbreaks.

5.4. Comparison of spatial accuracy

In addition to the comparison of detection times that was described above, we also computed the average spatial accuracy (the degree of overlap between true and detected clusters) for each method for each day of outbreak ($t = 1, \ldots, 14$). Results were averaged over all 2000 outbreaks. Letting S^* represent the detected region $S^* = \arg\max_S \{F(S)\}$, and S_T represent the true inject region (the subset of locations for which simulated cases were actually injected), the *overlap coefficient* is defined as Overlap $= \Sigma_{s_i \in S^* \cap S_T} w_i / \Sigma_{s_i \in S^* \cup S_T} w_i$, where w_i is the weight of location s_i. Similarly, we define Precision $= \Sigma_{s_i \in S^* \cap S_T} w_i / \Sigma_{s_i \in S^*} w_i$ and Recall $= \Sigma_{s_i \in S^* \cap S_T} w_i / \Sigma_{s_i \in S_T} w_i$. In each case, we define $w_i = \Sigma_t c_i^t / \Sigma_i \Sigma_t c_i^t$, where the sums are taken over all time steps and all locations. Each evaluation metric varies between 0 and 1, with Overlap $= 1$ if $S^* = S_T$, Precision $= 1$ if $S^* \subseteq S_T$ and Recall $= 1$ if $S_T \subseteq S^*$. Each spatial location was weighted proportionally to the total number of cases observed in that location, which can also be thought of as a proxy for the at-risk population. In our simulations, the expected number of cases injected into each affected location was also chosen proportionally to its weight. We believe that the weighted metrics are preferable to the unweighted metrics ($w_i = 1$ for all locations) since the total number of injected cases was small: many locations did not receive any injects on a given day, and some low weight locations (despite being considered part of the outbreak region) may not have received any injected cases over the entire duration of outbreak.

We first compared the overlap coefficient for each method on the last outbreak day ($t = 14$). The circular scan achieved an average overlap of 50.4% averaged over all 10 types of outbreak, whereas the all-subsets method had a much lower overlap of 32.0%. The fixed k fast localized scan achieved highest spatial accuracy for $k = 10$, with an overlap coefficient of 51.1%. The fixed r fast localized scan achieved highest accuracy for $r = 0.06$, with an overlap coefficient of 47.4%. The fast multiscan k method achieved highest accuracy for parameter $L = 0.5$, with an overlap coefficient of 49.8%, and the fast multiscan r method achieved highest accuracy for parameter $L = 100$, with an overlap coefficient of 41.7%.

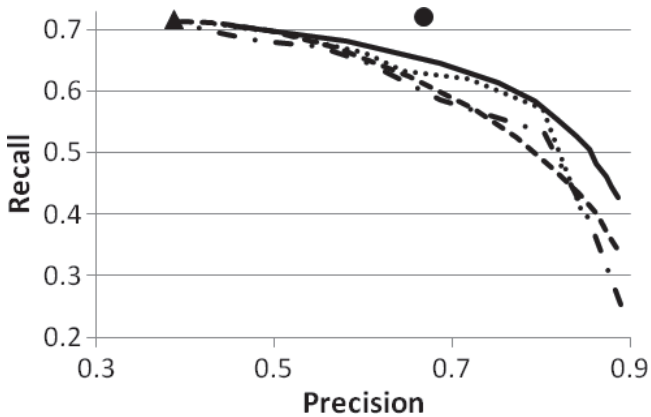

Fig. 6. Comparison of detection methods: trade-off between spatial precision and recall on the last day of outbreak for the fixed k (·····, $k = 5,\ldots,60$), fixed r (– · – ·, $r = 0.02,\ldots,0.24$), multiscan k (———, $L = 0.1,\ldots,1.2$) and multiscan r (– – –, $L = 20,\ldots,240$) methods, compared with the circles (●) and all-subsets (▲) approaches

Although these results suggest that the fast localized scan, fast multiscan and circular scan methods have similar spatial accuracy for well-chosen parameter values, in fact we see substantial differences in the size and shape of the regions detected. Fig. 6 shows the trade-off between spatial precision and recall for the fast localized scan and fast multiscan methods, compared with the circular scan and unconstrained subset scan. Increasing the neighbourhood size k or radius r for the fast localized scans, or decreasing the penalty L for the fast multiscans, tended to increase the size of clusters detected, thus increasing recall and decreasing precision. For the parameter values given above, the fast localized scan and fast multiscan methods had higher precision and lower recall than the circular scan. Also, as shown in Fig. 7, the relative performance of methods was highly dependent on the shape of outbreak: the fast localized scan and fast multiscan methods had higher precision and recall than the circular scan for elongated outbreaks, but lower recall for compact shapes of outbreak. This suggests that our methods tend to pick out the subset of locations that have been most affected by an outbreak, whereas the circular scan identifies a larger region which may contain both slightly affected and unaffected locations.

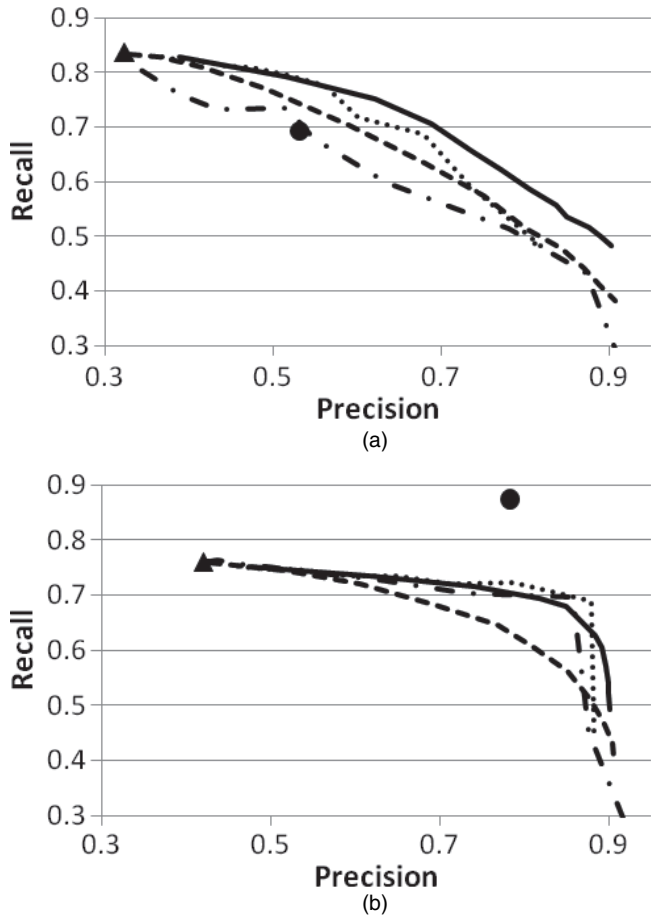

Fig. 7. Comparison of detection methods: trade-off between spatial precision and recall on the last day of outbreak for the fixed k (·····, $k = 5,\ldots,60$), fixed r (– · – ·, $r = 0.02,\ldots,0.24$), multiscan k (———, $L = 0.1,\ldots,1.2$) and multiscan r (– – –, $L = 20,\ldots,240$) methods, compared with the circles (●) and all-subsets (▲) approaches, on elongated clusters (5, 9 and 10) and compact clusters (1, 2 and 7): (a) elongated clusters; (b) compact clusters

6. Conclusions

This paper has presented three main contributions to the growing literature on event detection. First, we proposed a general framework for computationally efficient pattern detection based on the LTSS property, enabling rapid computation of the highest scoring subset of records in massive data sets. Although this paper focused primarily on pattern detection in univariate spatial and space–time data, LTSS can be applied to multivariate and non-spatial data sets as well. Second, we demonstrated that many commonly used spatial and space–time scan statistics satisfy the LTSS property, including Kulldorff's original spatial scan statistic and many recent variants including the EBP and EPG scan statistics. LTSS enables us to compute efficiently the highest scoring unconstrained subset of spatial locations for any of these statistics. However, an unconstrained search over subsets can return dispersed sets of locations that we would not consider to be 'spatial clusters' and typically underperforms the circular scan statistic for spatial event detection tasks. Thus we considered how spatial proximity constraints can be incorporated in the LTSS framework, either by placing hard constraints on the neighbourhood size or radius (in our fast localized scan methods) or by penalizing large neighbourhoods (in our fast multi-scan methods). We demonstrated both theoretically and empirically that these methods can efficiently maximize the likelihood ratio statistic subject to proximity constraints. Finally, our semisynthetic evaluation (using simulated respiratory outbreaks injected into real emergency department data) demonstrated that the resulting methods can substantially improve the power of detection, enabling 2 days faster detection of emerging outbreaks of disease with fewer than half as many missed outbreaks. Spatial accuracy (as measured by the degree of overlap between true and detected clusters) was improved for elongated and irregularly shaped outbreaks. For subtle outbreaks with a small number of injected cases, our methods tended to pick out the most affected zip codes, whereas the circular scan tended to identify a larger circular region which also contained slightly affected and unaffected zip codes.

Our current work focuses on extending the LTSS framework in several directions. First, although we have focused here on detecting patterns in univariate space–time data, LTSS can also be extended to the multivariate case, as discussed above. In Neill *et al.* (2010), we demonstrated that two variants of the multivariate space–time scan statistic (Burkom, 2003; Kulldorff *et al.*, 2007) can each be efficiently optimized over proximity-constrained subsets of locations and all subsets of the monitored data streams, even when the numbers of locations and streams are both very large. Similarly, in McFowland *et al.* (2011), we extend LTSS to general non-spatial data sets, efficiently optimizing a non-parametric scan statistic over subsets of records and attributes. This 'fast generalized subset scan' approach enables us to detect self-similar groups of data records which have anomalous values for some subset of attributes, with applications including customs monitoring (identifying patterns of illicit container shipments) and network intrusion detection (McFowland *et al.*, 2011).

Finally, LTSS can be used to accelerate spatial scans with other constraints, including shape constraints (maximizing the score function over all regions of a given shape) and connectivity constraints (maximizing the score function over all connected subgraphs). In each case, we have integrated LTSS into a 'branch-and-bound' framework, using the unconstrained all-subsets score of a group of locations as an upper bound on the constrained score, and ruling out subsets of locations which are provably non-optimal. We have recently developed GraphScan, a method for detection of arbitrary-shaped connected clusters in graph or network data (Speakman and Neill, 2010). GraphScan enables efficient, exact computation of the highest scoring connected clusters, with or without proximity constraints. Whereas Tango and Takahashi's flexible scan statistic (Tango and Takahashi, 2005) scales exponentially with neighbourhood size

and is computationally infeasible for neighbourhoods that are larger than 30 locations, GraphScan can easily scale up to over 100 locations, computing the highest scoring connected cluster in seconds. Similarly, in Neill (2008), we showed that LTSS can be used to scan over all distinct rectangular regions between 57 and 534 times faster than a naive search, requiring between 16 s and 2 min per day of data compared with over 2 h for a naive search. Unlike our original fast spatial scan method (Neill and Moore, 2004), which also searches over rectangular regions, the LTSS-enabled fast spatial scan can be used for any scan statistic satisfying the LTSS property (not just Kulldorff's statistic) and does not require locations to be mapped to a uniform grid.

Acknowledgements

This work was partially supported by the National Science Foundation, grants IIS-0916345, IIS-0911032 and IIS-0953330. A preliminary version was presented at the 2008 Annual Conference of the International Society for Disease Surveillance, and a one-page abstract was published in the journal *Advances in Disease Surveillance* (Neill, 2008).

References

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. New York: Wiley.

Buckeridge, D. L., Burkom, H. S., Moore, A. W., Pavlin, J. A., Cutchis, P. N. and Hogan, W. R. (2004) Evaluation of syndromic surveillance systems: development of an epidemic simulation model. *Morb. Mort. Wkly Rep.*, **53**, suppl., 137–143.

Burkom, H. S. (2003) Biosurveillance applying scan statistics with multiple, disparate data sources. *J. Urb. Hlth*, **80**, suppl. 1, i57–i65.

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

Duczmal, L., Cancado, A., Takahashi, R. and Bessegato, L. (2007) A genetic algorithm for irregularly shaped scan statistics. *Computnl Statist. Data Anal.*, **52**, 43–52.

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

Hogan, W. R., Cooper, G. F., Wallstrom, G. L., Wagner, M. M. and Depinay, J. M. (2007) The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by atmospheric release of Bacillus anthracis. *Statist. Med.*, **26**, 5225–5252.

Huang, L., Kulldorff, M. and Gregorio, D. (2007) A spatial scan statistic for survival data. *Biometrics*, **63**, 109–118.

Kulldorff, M. (1997) A spatial scan statistic. *Communs Statist. Theor. Meth.*, **26**, 1481–1496.

Kulldorff, M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic. *J. R. Statist. Soc.* A, **164**, 61–72.

Kulldorff, M., Athas, W., Feuer, E., Miller, B and Key, C. (1998) Evaluating cluster alarms: a space-time scan statistic and cluster alarms in Los Alamos. *Am. J. Publ. Hlth*, **88**, 1377–1380.

Kulldorff, M., Feuer, E. J., Miller, B. A. and Freedman, L. S. (1997) Breast cancer clusters in the northeast United States: a geographic analysis. *Am. J. Epidem.*, **146**, 161–170.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006) An elliptic spatial scan statistic. *Statist. Med.*, **25**, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K. and Platt, R. (2007) Multivariate scan statistics for disease surveillance. *Statist. Med.*, **26**, 1824–1833.

Kulldorff, M. and Nagarwalla, N. (1995) Spatial disease clusters: detection and inference. *Statist. Med.*, **14**, 799–810.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J. and Glance, N. (2007) Cost-effective outbreak detection in networks. In *Proc. 13th Int. Conf. Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery Press.

McFowland III, E., Speakman, S. and Neill, D. B. (2011) Fast generalized subset scan for anomalous pattern detection. *Technical Report*. Carnegie Mellon University, Pittsburgh.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R. and Kulasekera, V. (2003) Dead bird clustering: a potential early warning system for West Nile virus activity. *Emergng Infect. Dis.*, **9**, 641–646.

Neill, D. B. (2006) Detection of spatial and spatio-temporal clusters. *PhD Thesis*. School of Computer Science, Carnegie Mellon University, Pittsburgh.

Neill, D. B. (2008) Fast and flexible outbreak detection by linear-time subset scanning. *Adv. Dis. Surveill.*, **5**, 48.

Neill, D. B. (2009) An empirical comparison of spatial scan statistics for outbreak detection. *Int. J. Hlth Geograph.*, **8**, 20.
Neill, D. B. and Cooper, G. F. (2010) A multivariate Bayesian scan statistic for early event detection and characterization. *Mach. Learn.*, **79**, 261–282.
Neill, D. B. and Lingwall, J. (2007) A nonparametric scan statistic for multivariate disease surveillance. *Adv. Dis. Surveill.*, **4**, 106.
Neill, D. B., McFowland III, E. and Zheng, H. (2010) Fast subset scan for multivariate event detection. *Technical Report*. Carnegie Mellon University, Pittsburgh.
Neill, D. B. and Moore, A. W. (2004) Rapid detection of significant spatial clusters. In *Proc. 10th Association for Computing Machinery Conf. Knowledge Discovery and Data Mining*, pp. 256–265. New York: Association for Computing Machinery Press.
Neill, D. B., Moore, A. W., Sabhnani, M. R. and Daniel, K. (2005) Detection of emerging space-time clusters. In *Proc. 11th Association for Computing Machinery Conf. Knowledge Discovery and Data Mining*, pp. 218–227. New York: Association for Computing Machinery Press.
Nemhauser, G., Wolsey, L. and Fisher, M. (1978) An analysis of the approximations for maximizing submodular set functions. *Math. Program.*, **14**, 265–294.
Patil, G. P. and Taillie, C. (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ. Ecol. Statist.*, **11**, 183–197.
Speakman, S. and Neill, D. B. (2010) Fast graph scan for scalable detection of arbitrary connected clusters. In *Proc. 2009 International Society for Disease Surveillance A. Conf.* International Society for Disease Surveillance.
Tango, T. and Takahashi, K. (2005) A flexibly shaped spatial scan statistic for detecting clusters. *Int. J. Hlth Geog.*, **4**, 11.
Wallstrom, G. L., Wagner, M. M. and Hogan, W. R. (2005) High-fidelity injection detectability experiments: a tool for evaluation of syndromic surveillance systems. *Morb. Mort. Wkly Rep.*, **54**, suppl., 85–91.