Computational Statistics and Data Analysis 157 (2021) 107149

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Support vector subset scan for spatial pattern detection

Dylan Fitzpatrick [a,*], Yun Ni [a], Daniel B. Neill [b]

[a] Carnegie Mellon University, Pittsburgh, PA, United States of America
[b] New York University, New York, NY, United States of America

ARTICLE INFO

Article history:
Received 9 March 2020
Received in revised form 2 November 2020
Accepted 15 November 2020
Available online 13 December 2020

Keywords:
Anomalous pattern detection
Machine learning
Spatial analysis
Subset scanning

ABSTRACT

Discovery of localized and irregularly shaped anomalous patterns in spatial data provides useful context for operational decisions across many policy domains. The support vector subset scan (SVSS) integrates the penalized fast subset scan with a kernel support vector machine classifier to accurately detect spatial clusters without imposing hard constraints on the shape or size of the pattern. The method iterates between (1) efficiently maximizing a penalized log-likelihood ratio over subsets of locations to obtain an anomalous pattern, and (2) learning a high-dimensional decision boundary between locations included in and excluded from the anomalous subset. On each iteration, location-specific penalties to the log-likelihood ratio are assigned according to distance to the decision boundary, encouraging patterns which are spatially compact but potentially highly irregular in shape. SVSS outperforms competing methods for spatial cluster detection at the task of detecting randomly generated patterns in simulated experiments. SVSS enables discovery of practically-useful anomalous patterns for disease surveillance in Chicago, IL, crime hotspot detection in Portland, OR, and pothole cluster detection in Pittsburgh, PA, as demonstrated by experiments using publicly available data sets from these domains.

1. Introduction

Detecting anomalous patterns in spatial data has applications across a wide variety of policy domains. Public health agencies may be interested in characterizing spatial regions with high prevalence of disease, indicating a possible outbreak. In large cities, police analysts are interested in detecting and characterizing flare-ups of violent crime in order to dispatch patrols effectively. Identifying spatial clusters of citizen complaints can help agencies responsible for city services such as road maintenance or sanitation to prioritize projects and efficiently address complaints. In this paper, we present an approach to address such examples where decision makers must identify spatial patterns to design and target policy interventions. In real world settings,[1] we may expect patterns to be highly irregular in shape, as spatial clustering is often influenced by environmental or social factors such as transportation patterns, built infrastructure, land use, or natural features. Our proposed method allows for precise localization of spatial clusters regardless of shape, addressing the need for flexible detection of spatial clusters for intervention.

Anomalous patterns which are *spatially compact* are often preferable for identifying situations in need of intervention or for guiding operational decisions in policy applications. An anomalous cluster of locations is spatially compact if member

* Correspondence to: 4800 Forbes Avenue, Pittsburgh, PA, United States of America.
E-mail address: djfitzpa@alumni.cmu.edu (D. Fitzpatrick).

[1] For ease of reproducibility, the three real-world data sets described in this work are all publicly available and processed versions of the data sets are posted as supplementary material alongside code as an annex to this work.

https://doi.org/10.1016/j.csda.2020.107149

locations are situated close to each other in space and non-anomalous locations are sparse within the boundaries of the cluster. Spatially compact clusters may be preferable for targeting interventions because they are more likely to correspond to a single structural cause (e.g., virus-carrying mosquitoes breeding in a pool of stagnant water), or because locations in these clusters can be efficiently targeted for mitigation efforts due to their physical proximity (e.g., a cluster of potholes on a highly-trafficked road can be repaired by a single maintenance crew). Yet simpler *spatial scan* approaches such as Kulldorff (1997), which search for spatially compact clusters of a fixed shape, may fail to correctly identify the spatial extent of the cluster, and have reduced detection power when the cluster is elongated or irregular in shape.

This work builds on the *subset scan* approach to pattern detection, which finds anomalous patterns by performing a constrained scan over subsets of data points. In this framework, the *anomalousness* of fixed subsets can be evaluated using a predefined score function, such as the log-likelihood ratio statistics applied in Kulldorff (1997), Neill et al. (2005), Neill (2009), and Neill (2012). The subset scan approach has demonstrated high power to detect both localized and global patterns, unlike 'bottom-up' approaches which identify and aggregate individual anomalies, and 'top-down' approaches which localize anomalous patterns detected in aggregated data (Neill, 2009, 2012). Outlier detection methods such as one-class SVM (Schölkopf et al., 2001) are likely to pick out individually anomalous data records with high counts (often due to chance) and thus fail to detect the regions of interest, while density-based clustering methods such as DBSCAN (Ester et al., 1996) can find anomalous regions but are dramatically outperformed by our proposed method (as shown in Section 3.1).

Subset scanning poses a significant computational challenge, as there exist 2^N possible subsets to consider when searching for the most anomalous subset for a data set containing N elements. Several approaches have been proposed to reduce the computation needed to search over the entire data set. One approach is to restrict the search space by considering only regions of a specific shape, such as circles (Kulldorff, 1997), ellipses (Kulldorff et al., 2006), or rectangles (Neill and Moore, 2004). Other approaches reduce the number of subsets under consideration by enforcing connectivity constraints between elements included in a subset (Patil and Taillie, 2004; Duczmal and Assuncao, 2004; Duczmal et al., 2007; Yiannakoulias et al., 2007; Takahashi et al., 2008; Costa and Kulldorff, 2014; Speakman et al., 2015). These methods enable efficient computation of anomalous patterns but sacrifice both detection power and spatial accuracy in comparison to unconstrained methods which do not restrict the search space (Neill, 2012).

One alternative to subset scanning for anomalous pattern detection is to fully model dependence across spatially-distributed point observations using a geostatistical model, such as a spatial generalized linear mixed model (SGLMM). Introduced in Diggle et al. (2002), SGLMMs are a form of generalized linear model in which spatial dependence is modeled with Gaussian processes across the spatial extent. While SGLMMs represent a powerful tool for modeling spatial data, standard sample-based inference approaches on these models are computationally expensive and slow to converge (Haran, 2011). Further, SGLMMs do not define a decision boundary around anomalous spatial regions, which is a practically useful output for characterizing the extent of an affected region and enabling targeted interventions.

Identifying patterns with arbitrary shape in the subset scanning framework is non-trivial given the high number of patterns to consider. Methods that search over subsets with a fixed geometric shape or impose connectivity constraints are not likely to accurately characterize affected regions with irregular shapes or multiple disconnected components. Connectivity may also be difficult to determine in contexts where no inherent graph structure is obvious. Underconstrained patterns which are too disconnected or sparse may be similarly unrealistic. Thus, recent developments in spatial scanning have focused on encouraging patterns which are spatially compact while still allowing for detection of irregular shapes. Duczmal et al. (2006) and Yiannakoulias et al. (2007) propose penalized score functions that discourage highly irregular shapes based on measures of non-compactness or non-connectivity, but do not provide a statistical framework for interpreting the penalized versions of the score functions. Other approaches have applied multi-objective optimization algorithms to simultaneously maximize a score function and minimize a geometric penalty function (Duarte et al., 2010; Cancado et al., 2010; Duczmal et al., 2012; Moreira et al., 2015). These multi-objective methods result in a set of non-dominated candidate patterns which must then be ranked by a single objective function to obtain the most anomalous pattern. This ranking step presents both computational and theoretical difficulties, as the set of candidates may be large and the desired tradeoff between multiple objectives could be ill-defined across candidate patterns.

Neill (2012) presents the fast subset scan (FSS), demonstrating that the most anomalous unconstrained subset across an entire data set can be found both efficiently and exactly for a family of score functions satisfying the Linear Time Subset Scan property. In practice, the FSS framework may detect patterns which are spread across the spatial extent of the study area and sparsely distributed among non-anomalous points. Several approaches to imposing hard spatial constraints on FSS have been proposed, such as searching only over local neighborhoods consisting of each location and its $k-1$ neighbors (Neill, 2012), or searching over locations connected by an underlying graph structure (Speakman et al., 2015). Speakman et al. (2016) provide a structured approach to incorporating soft constraints into the FSS framework with the penalized fast subset scan (PFSS), showing that one can apply additive penalties and still maximize the penalized score function efficiently and exactly. While PFSS gives us a framework for incorporating soft constraints, the question of how to define penalty terms to encourage spatial compactness in detected patterns remains open.

In this work, we present the support vector subset scan (SVSS), which detects anomalous patterns in spatial data that are spatially coherent but potentially highly irregular in shape. SVSS integrates PFSS with a kernel support vector machine (SVM) to encourage compact subsets of locations. The SVM provides a natural solution to the problem of specifying element-specific penalties for PFSS such that detected patterns are geometrically compact but unconstrained in size, shape, or connectivity. SVSS benefits from the ability of PFSS to detect subtle but significant anomalous patterns, while

leveraging the SVM to identify coherent spatial regions with a high density of anomalous points. This novel combination of two proven methods results in a new approach for anomalous pattern detection that outperforms each of the individual component methods. SVSS imposes soft constraints on FSS, which encourage spatial compactness at the cost of a lower anomalousness score. In comparison to the sparse patterns returned by FSS, SVSS finds compact patterns that are more suitable for targeted intervention.

The SVSS algorithm proceeds iteratively, alternating between efficiently maximizing a penalized log-likelihood ratio (LLR) over subsets of locations, and learning a high-dimensional decision boundary between locations included in and excluded from the anomalous subset. Location-specific penalties are computed according to distance to the decision boundary and added to the LLR score function, resulting in anomalous patterns which are spatially compact and irregular in shape. This iterative method is guaranteed to converge to a locally-optimal subset with respect to the biconvex SVSS objective function (Gorski et al., 2007). We apply multiple random restarts to approach the global optimum of the SVSS objective.

In Section 2, we provide the statistical background motivating SVSS, then define the SVSS optimization problem and algorithm. In Section 3.1, we evaluate SVSS on the task of detecting letter-shaped anomalous patterns in simulated data and find that it significantly outperforms competing methods at finding patterns which closely approximate the true affected region. In Sections 3.2–3.4, we demonstrate the method in three real world contexts where spatial pattern detection is useful for guiding operations and policy decisions using publicly available data sets. In the domain of disease surveillance, we apply SVSS to West Nile Virus test results to identify disease clusters throughout the city of Chicago, IL. For crime surveillance, we apply SVSS to characterize hotspots of street crime in Portland OR. Finally, we apply SVSS to detect clusters of potholes in Pittsburgh, PA, demonstrating the utility of the method for city services and management. We end with concluding remarks in Section 4.

2. Support vector subset scan (SVSS)

In this section, we describe a parametric scan statistic approach for spatial pattern detection under weak distributional assumptions.

2.1. Background: Penalized fast subset scan (PFSS)

Consider the setting in which data set D includes a set of spatial coordinates $\mathbf{x}_i$ for locations ($i = 1, \ldots, N$). Let $\boldsymbol{\alpha} \in \{0, 1\}^N$ be a vector specifying a subset of locations, with $\alpha_i = 1$ if location i is included in the subset and $\alpha_i = 0$ otherwise. Maximizing a score function over subsets is performed by searching over values of the vector $\boldsymbol{\alpha}$ and maximizing some score function $F(\boldsymbol{\alpha})$. Neill et al. (2005) propose a class of score functions called *expectation-based scan statistics*, in which data set D also includes observed values (or "counts") $\mathbf{c}$ and expected values (or "baselines") $\mathbf{b}$ of a random field indexed at locations. These location-specific observations and expected values provide the basis for defining $F(\boldsymbol{\alpha})$ and determining whether a subset is anomalous.

Let $H_1(\boldsymbol{\alpha})$ be an alternative hypothesis that assumes an event occurring in the subset defined by $\boldsymbol{\alpha}$ causing increased values at those locations, and let H_0 be the null hypothesis that assumes no event occurring in the data set (or equivalently, that $\alpha_i = 0$ for all i). Following Kulldorff (1997) and Neill et al. (2005), we define our score function as a log-likelihood ratio (LLR) statistic $F(\boldsymbol{\alpha}) = \log(Pr(D|H_1(\boldsymbol{\alpha}))/Pr(D|H_0))$. The expectation-based scan statistics assume that under alternative hypothesis $H_1(\boldsymbol{\alpha})$, values c_i are drawn with mean qb_i inside of the region defined by $\boldsymbol{\alpha}$ and mean b_i outside of that region for some multiplicative constant factor $q > 1$ known as *relative risk*. The expectation-based scan statistic is formulated as

$$F(\boldsymbol{\alpha}) = \max_{q>1} \sum_{i=1}^{N} \alpha_i \big[\log Pr(c_i|qb_i) - \log Pr(c_i|b_i)\big]. \tag{1}$$

Speakman et al. (2016) introduce the Penalized Fast Subset Scan (PFSS), observing that for a fixed value of relative risk q, the LLR for the exponential family of expectation-based scan statistics can be expressed as an additive set function over all locations included in a subset:

$$F(\boldsymbol{\alpha}|q) = \sum_{i=1}^{N} \alpha_i \lambda_i(q),$$

and

$$\begin{aligned} F(\boldsymbol{\alpha}) &= \max_{q>1} F(\boldsymbol{\alpha}|q), \\ &= \max_{q>1} \sum_{i=1}^{N} \alpha_i \lambda_i(q), \end{aligned}$$

where λ_i terms depend only on observed values c_i, baselines b_i, and relative risk q. Because $\lambda_i(q)$ expressions can be derived for a variety of expectation-based scan statistics, this additive score function is flexible in the underlying data

distribution, making the assumption that observed values c_i are drawn from a distribution in the exponential family with finite first moments. Further, the additive property of the score function enables addition of location-specific penalty terms to the LLR, denoted as Δ_i. Each Δ_i can be interpreted as the prior log-odds that location i is included in the affected subset. We express the penalized score function as

$$F_{pen}(\boldsymbol{\alpha}) = \max_{q>1} \sum_{i=1}^{N} \alpha_i \big(\lambda_i(q) + \Delta_i\big), \tag{2}$$

where $\lambda_i(q) + \Delta_i$ represents the total contribution of location i to the score function. Conditioning on a fixed value of relative risk q, the penalized score function can be optimized over all subsets by including all and only those locations with a positive total contribution $\lambda_i(q) + \Delta_i$. The score functions for expectation-based scan statistics can be optimized efficiently by considering at most $2N$ distinct values of q (Speakman et al., 2016).

PFSS thus provides an extremely flexible and computationally efficient framework for scanning over subsets and incorporating soft constraints to encourage patterns with desirable attributes. Yet the PFSS framework is not sufficient to obtain spatial coherence or compactness in detected patterns. Because the penalty terms Δ_i must be decided for each element before optimizing the penalized LLR, there is no natural way of assigning element-wise bonuses or penalties if we do not already know where the anomalous subset is likely to be. PFSS is also limited because the penalties must be location-specific, precluding application of an arbitrary prior over subsets to encourage more coherent regions. If the penalties depended on other locations in the subset, we would not be able to perform the scan efficiently and would have to exhaustively enumerate subsets. Thus, we must incorporate additional tools in order to specify location-specific Δ_i terms which promote compactness.

2.2. Background: Support vector machines

To formulate Δ_i terms for the PFSS framework, we turn to the support vector machine (SVM), a popular algorithm for binary classification first proposed by Cortes and Vapnik (1995). The SVM is trained on data elements consisting of feature vectors $\mathbf{x}_i$ and positive or negative class labels y_i, finding the separating hyperplane between classes which maximizes the margin between classes, or the distance between the hyperplane and the nearest data point on either side.

A soft-margin SVM introduces slack variables ξ_i and tuning parameter C to address the case when the two classes are not linearly separable. Learning an SVM is formulated as the following optimization problem, where weight vector $\mathbf{w}$ and intercept term w_0 define the separating hyperplane and ϕ is a nonlinear transformation which maps $\mathbf{x}$ to a high-dimensional feature space and allows a nonlinear decision boundary in the original space:

$$\begin{aligned} &\min_{\xi,\mathbf{w},w_0} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i, \\ &\text{subject to } \xi_i \geq 0, \forall i = 1,\ldots,N, \\ &\qquad \text{and } y_i(\mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0) \geq 1 - \xi_i, \forall i = 1,\ldots,N. \end{aligned}$$

The SVM with Gaussian kernel results in a decision boundary with a potentially highly irregular shape and multiple disconnected components, tending to demarcate high-density regions of each class in the original feature space. As a supervised method, the SVM requires locations to be labeled as belonging to one class or the other. In the context of spatial pattern detection, it is not obvious how to assign labels for the SVM. One straightforward approach is to apply a threshold to some function of location-specific counts and baselines to assign class labels (included for comparison in Section 3.1). However, particularly for subtle signals, a high proportion of points will initially be mislabeled, leading to an extremely noisy classification problem and resulting poor performance. These considerations motivate our SVSS approach which alternates between PFSS and SVM optimization steps: we iteratively pick good thresholds for class labeling using the penalized score function from PFSS, then specify location-specific penalty terms as distances to a SVM hyperplane within the PFSS framework. The resulting patterns are spatially coherent but irregular in shape due to the nonlinear decision boundary given by the kernel SVM.

2.3. SVSS optimization problem

With the kernel SVM, we now have the tools necessary to specify Δ_i penalty terms for PFSS in the context of unsupervised subset scanning. Given a fixed $\boldsymbol{\alpha}$ which defines a subset of locations, let $y_i = 2\alpha_i - 1$ for all locations. Thus, our class labels $y_i \in \{-1, 1\}$ represent inclusion or exclusion from the given subset defined by $\boldsymbol{\alpha}$, so that the SVM learns a decision boundary to separate included from excluded locations. We formulate SVSS as a modified version of the SVM optimization problem, while also minimizing over subsets $\boldsymbol{\alpha}$ and including the unpenalized LLR score function $F(\boldsymbol{\alpha})$ as an additional regularization term. Alternatively, we could view this as a maximization of the penalized scan statistic,

optimizing $F(\boldsymbol{\alpha})$ with penalties from the SVM slack variables and the width of the margin. We now have two tuning parameters C_0 and C_1, controlling the relative importance of these three factors.

$$\begin{aligned}\min_{\boldsymbol{\alpha},\xi,\mathbf{w},w_0} \frac{1}{2}\|\mathbf{w}\|^2 + C_0\sum_{i=1}^{N}\xi_i - C_1F(\boldsymbol{\alpha}),\\ \text{s.t. } \alpha_i \in \{0,1\}, \forall i = 1,\ldots,N,\\ \xi_i \geq 0, \forall i = 1,\ldots,N,\\ \text{and } (2\alpha_i - 1)(\mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0) \geq 1 - \xi_i, \forall i = 1,\ldots,N.\end{aligned}$$

Equivalently, we can express slack variables as a function of α_i to obtain the SVSS optimization problem.

$$\begin{aligned}\min_{\boldsymbol{\alpha},\xi,\mathbf{w},w_0} \frac{1}{2}\|\mathbf{w}\|^2 + C_0\sum_{i=1}^{N}\xi_i(\alpha_i) - C_1F(\boldsymbol{\alpha}),\\ \text{s.t. } \alpha_i \in \{0,1\}, \forall i = 1,\ldots,N,\\ \text{and } \xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0)).\end{aligned} \tag{3}$$

This optimization problem is not convex, making computation of the global optimum non-trivial. We optimize the SVSS objective by alternately (1) fixing the anomalous subset $\boldsymbol{\alpha}$ and optimizing $\mathbf{w}$ and w_0 by training the SVM, then (2) fixing $\mathbf{w}$ and w_0 and learning an optimal subset $\boldsymbol{\alpha}$ through a search over subsets to maximize the score function. This alternating approach to minimization is guaranteed to give a convergent sequence for the biconvex SVSS objective function, but does not necessarily find the global optimum (Gorski et al., 2007). Thus, we use multiple restarts to randomly initialize the Δ_i penalty terms and take the optimal subset across all restarts (as measured by the combined objective) as our anomalous pattern. SVSS iterates over two computationally efficient algorithms (PFSS and SVM). Each iteration of PFSS is an $O(N \log N)$ operation. Computational complexity of the RBF-kernel SVM scales between $O(N^2)$ and $O(N^3)$ and is dependent on the specific data set and amount of regularization applied (Bottou and Lin, 2007). In practice, across a variety of data sets, only a small number of iterations are needed for the algorithm to converge to a local optimum. Algorithm 1 outlines the SVSS algorithm using T_{max} random restarts.

Algorithm 1 Support Vector Subset Scan

procedure SVSS($\mathbf{c}$, $\mathbf{b}$, $\mathbf{x}$, T_{max}, C_0, C_1) ▷ Values $\mathbf{c}$, expectations $\mathbf{b}$, and coordinates $\mathbf{x}$
  $min\_score \leftarrow \infty$
  for $t := 1$ **to** T_{max} **do** ▷ T_{max} random restarts
    $\xi_i(\alpha_i) \leftarrow \text{Uniform}(-C_0, C_0), \forall i = 1, ..., N$
    while $\boldsymbol{\alpha}$ is changing **do**
      $\boldsymbol{\alpha} \leftarrow \underset{\boldsymbol{\alpha}}{\text{argmax}}\ F(\boldsymbol{\alpha}) - (C_0/C_1)\sum_{i=1}^{N}\xi_i(\alpha_i)$ ▷ Optimize over $\boldsymbol{\alpha}$
      $\boldsymbol{\xi}, \mathbf{w}, w_0 \leftarrow \underset{\boldsymbol{\xi},\mathbf{w},w_0}{\text{argmin}} \frac{1}{2}||\mathbf{w}||^2 + C_0\sum_{i=1}^{N}\xi_i(\alpha_i)$ ▷ Optimize over $\mathbf{w}$, w_0
    end while
    $score \leftarrow \frac{1}{2}||\mathbf{w}||^2 + C_0\sum_{i=1}^{N}\xi_i(\alpha_i) - C_1F(\boldsymbol{\alpha})$
    if $score < min\_score$ **then**
      $min\_score \leftarrow score$
      $\boldsymbol{\alpha}_{min} \leftarrow \boldsymbol{\alpha}$
    end if
  end for
  return $\boldsymbol{\alpha}_{min}$
end procedure

For a given hyperplane specified by a fixed $\mathbf{w}$ and w_0, we can optimize the SVSS objective using the PFSS algorithm. Optimizing the SVSS objective for fixed $\mathbf{w}$ and w_0 is equivalent to

$$\underset{\boldsymbol{\alpha}}{\text{argmax}}\ F(\boldsymbol{\alpha}) - \frac{C_0}{C_1}\sum_{i=1}^{N}\xi_i(\alpha_i),$$

where

$$\xi_i(\alpha_i) = \begin{cases}\max(0, 1 - \mathbf{w}\cdot\phi(\mathbf{x}_i) + w_0), & 2\alpha_i - 1 = +1\\ \max(0, 1 + \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0), & 2\alpha_i - 1 = -1.\end{cases}$$

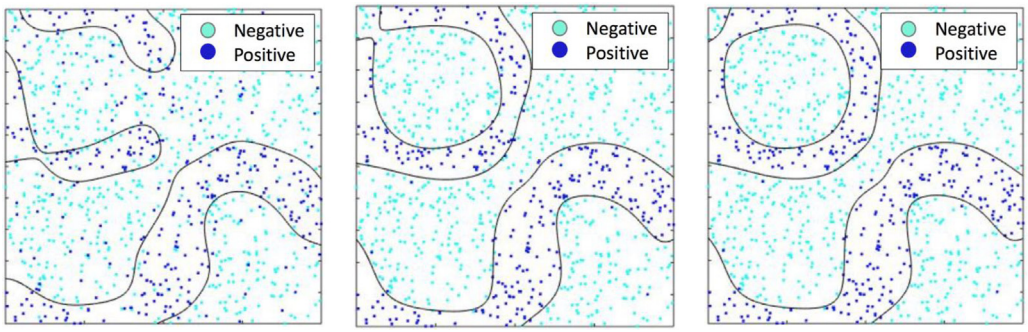

Fig. 1. Refinement of the detected pattern (shown in dark blue) across iterations of SVSS. On the left, the pattern detected by the first iteration of the Fast Subset Scan includes many points outside the true affected region. In the second (middle) and third (right) iterations, points outside the SVM decision boundary are penalized and the detected pattern improves, rapidly approaching the true affected region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Without changing the optimal solution, we can solve a modified problem with penalty terms that are non-zero only for points included in the subset defined by a fixed $\boldsymbol{\alpha}$:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, F(\boldsymbol{\alpha}) - \frac{C_0}{C_1}\sum_{i=1}^{N}\alpha_i\Delta_i, \tag{4}$$

where

$$\begin{aligned}\Delta_i &= \max(0, 1 - \mathbf{w}\cdot\phi(\mathbf{x}_i) + w_0) - \max(0, 1 + \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0),\\ &= \begin{cases} \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0 + 1, & \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0 \geq 1\\ 2(\mathbf{w}\cdot\phi(\mathbf{x}_i) - = w_0), & \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0 \in (-1, 1)\\ \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0 - 1, & \mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0 \leq -1. \end{cases}\end{aligned}$$

Because each penalty term Δ_i depends only on spatial coordinates from location i and not other locations, we can efficiently optimize (4) using the PFSS algorithm. Specifically, for a fixed relative risk q, we include only those locations with a positive total contribution to the objective function, and we maximize the objective over linearly many values of q as discussed in Section 2.1.

Refinement of the detected pattern across iterations of the SVSS algorithm is demonstrated in Fig. 1. As the algorithm progresses, points outside of the SVM decision boundary are penalized, resulting in patterns with spatial coherence. Fig. 2 shows the values of the penalty term Δ generated by SVSS on the final iteration of the algorithm across the spatial region surrounding a simulated anomalous pattern. Fig. 3 shows the patterns returned by SVSS and circular scanning windows in the presence of both hot spots (with increased counts relative to baseline) and cold spots (with decreased counts). The presence of a cold spot contained *within* the hot spot does not affect the ability of SVSS to detect the surrounding hot spot, but the cold spot forces the circular scan to identify only a small portion of the true hot spot. While this work focuses on applying SVSS for detecting hot spots with elevated values, the method can also be applied for cold spot detection with a minor change to the log-likelihood ratio specification. We note that, because of the flexibility of the SVSS approach in identifying irregularly-shaped regions, we would typically expect little or no overlap between the areas returned from hot-spot and cold-spot detection. In contrast, if the intensity of the cold spot is reduced in Fig. 3(a), the circular scan identifies concentric circles with the cold spot included as part of the hot spot. If SVSS does identify a non-disjoint hot spot and cold spot (e.g., in the case of two intersecting lines, one "hot" and one "cold"), the overlap area could be considered part of both hot spots, or neither, at the user's discretion.

2.4. Ranking disconnected regions

As previously noted, the decision boundary learned by an SVM may result in multiple disconnected components, allowing the SVSS algorithm to return anomalous patterns with multiple disjoint regions. In the subset scanning framework, it is reasonable to consider these regions as a single anomalous pattern, because the problem formulation assumes a constant relative risk q across the entire pattern. However, for some applications, we may seek to further search over components of our pattern to find the most anomalous component across disconnected regions of our pattern. To accomplish this, we optimize the penalized LLR $F_{pen}(\boldsymbol{\alpha})$ over components of the final SVM decision boundary as a post-processing step. We can also consider the convex hull of a connected component in the grid in order to evaluate geometric characteristics of the region, such as the compactness measure discussed in Section 3. We demonstrate this ranking approach to connected components in Sections 3.2 and 3.3 .

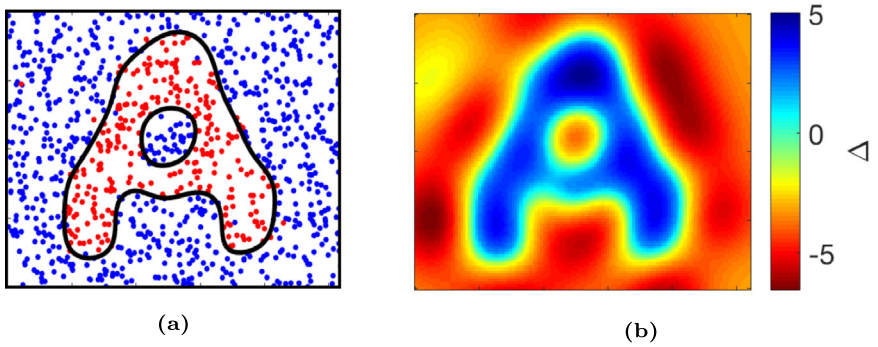

Fig. 2. (a) Binary classification with kernel SVM on final iteration of SVSS. SVM decision boundary shown in black. (b) Penalty surface learned on final iteration based on distance to separating hyperplane.

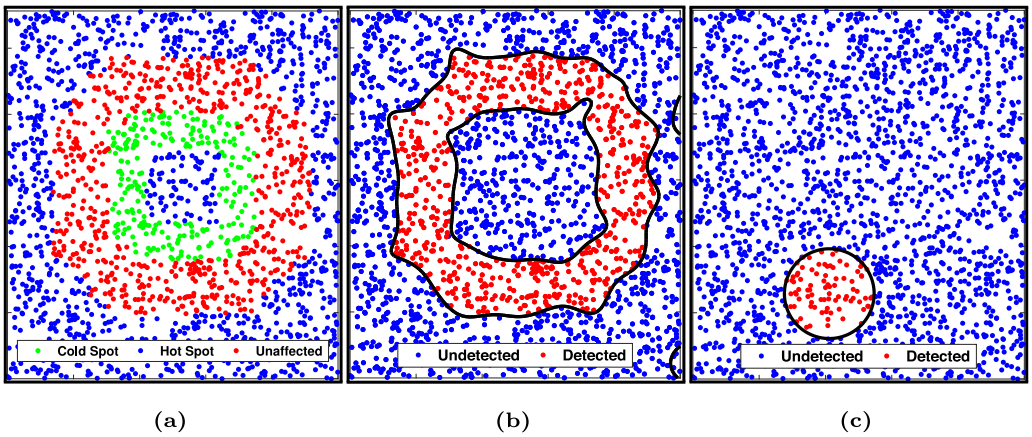

Fig. 3. Anomalous patterns detected by two methods in the presence of both hot spots and cold spots. (a) True labels of spatial locations, with hot spots shown in red and cold spots shown in green. (b) Pattern detected by SVSS. (c) Pattern detected by the circular scan. The presence of a cold spot forces the circular scan to identify only a portion of the true affected region, while SVSS is able to closely approximate the spatial extent of the affected region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

After running SVSS to obtain an optimal pattern, we take the following approach to find disjoint components of the pattern. We first overlay a grid of equally-spaced points over our spatial extent, and classify points using the SVM classifier learned on the final iteration of SVSS. We then find the connected components within the grid belonging to the positive (anomalous) class. Locations in the optimal pattern detected by SVSS are assigned to the connected component of the nearest point from the grid overlay. The resolution of the grid overlay is selected such that any disjoint components separated by less distance than the grid resolution can be practically considered a single pattern component.

2.5. Tuning parameters

The SVSS optimization problem includes several parameters which must be selected ahead of time. C_0 is a regularization parameter which controls the impact of misclassification on the overall objective function during the SVM step. With higher C_0, the SVM learns a more complex decision boundary to avoid misclassifications, giving patterns that are more irregular in shape. Similarly, the kernel function chosen for the SVM step may have a tuning parameter which affects the shape of the decision boundary such as the bandwidth parameter for a Gaussian kernel. C_1 should be chosen in relation to the value of C_0, as the ratio $\frac{C_0}{C_1}$ controls the scale of the penalty terms relative to the LLR in the PFSS.

In practice, the choice of parameters can have a significant impact on the shape and size of the patterns returned by SVSS. High values of C_0 and low values of the Gaussian kernel bandwidth parameter can result in highly irregular and elongated patterns that likely capture noise in the data rather than true anomalous patterns. A procedure is needed for selecting parameter values that avoids overfitting to noise in the data while still enabling SVSS to capture truly irregular affected regions. To tune the SVSS parameters, we perform 10-fold cross-validation and choose the set of parameters that results in the highest average anomalousness score on held-out data. Specifically, we choose the parameters which maximize the average *unpenalized* LLR for points classified as anomalous by the SVM trained in the final iteration of SVSS. The unpenalized LLR on held-out data corresponds to how well the identified SVSS decision boundary (for

particular parameter settings) captures the latent risk surface. By optimizing on multiple held-out data folds, we prevent overfitting to noise, and the complexity of the resulting patterns reflects the true underlying spatial distribution. For a fixed data set, we observe minimal variation in the optimal values of C_0 and C_1 selected across multiple random restarts, providing evidence that the optimal parameter choices are a function of patterns in the true underlying data distribution. We therefore reduce computation time by completing the parameter tuning step once rather than tuning parameters separately for each restart.

3. Evaluation and results

Evaluation of pattern detection methods on real world data can be difficult, given that we often do not know the true affected region that we hope to capture with the detected patterns. We evaluate the performance of SVSS and other pattern detection methods on simulated experiments where ground truth is known, then demonstrate SVSS in three real world pattern detection settings using real data.

3.1. Detecting letter-shaped simulated patterns

To evaluate our method in an experimental setting, we generate patterns of varying size, shape, and intensity in simulated data. On each run of the simulation, we draw 2000 locations uniformly at random across a rectangular study area. To generate patterns of irregular shape, we insert an *affected region* within the study area with shape matching a letter from the English alphabet. Each location has an observed count c_i drawn from the Poisson distribution, with counts outside the affected region drawn $c_i \sim Poisson(100)$ and counts inside the region drawn $c_i \sim Poisson(100 + intensity)$. Each location has a fixed baseline $b_i = 100$. We report average performance across 1300 simulations (50 simulations for each of the 26 letters in the uppercase English alphabet) for each pattern size under consideration, ranging from 1% to 20% of the study area. We tune parameters for SVSS using the cross-validation procedure outlined in Section 2.5. For all data sets considered, we observe minimal variation in the subsets returned and the LLR of optimal subsets across multiple, randomly initialized restarts for SVSS, and we therefore fix the number of restarts at 10 for all experiments. We compare the performance of SVSS with five other methods for spatial pattern detection: the *circular scan statistic* (Kulldorff, 1997), *upper level set scan statistic (ULS)* (Patil and Taillie, 2004), the *fast subset scan* (FSS) (Neill, 2012), *DBSCAN* with thresholding (Ester et al., 1996), and the *Kernel Support Vector Machine* (kSVM) with thresholding. Implementation details for these methods are included in Appendix A. All experiments were run in MATLAB R2016a.

For the circular scan, ULS, FSS, and SVSS, we apply the expectation-based Poisson scan statistic to formulate the LLR. We evaluate the performance of all methods at capturing the true affected region with the top pattern returned using *precision* and *recall*. Precision is defined as the proportion of points in the top pattern that lie in the true affected region, or the number of true positives divided by the number of true and false positives. Recall (or true positive rate) is defined as the proportion of points in the true affected region that are included in the top pattern, or the number of true positives divided by the number of true positives and false negatives.

We first report summary statistics from the six pattern detection methods on individual samples from three different signal intensities. For a pattern S returned by one of the scanning algorithms, we report the number of locations included in the pattern (n_S) and two measures of anomalousness: the unpenalized log-likelihood ratio (*LLR*), and the maximum likelihood estimate of the relative risk q_{MLE}. We also adopt a measure of geometric compactness presented in Duczmal et al. (2006). For a zone z, the geometric compactness $K(z)$ is defined as the area of z divided by the area of the circle with the same perimeter as the convex hull of z. This measure of compactness is highest for circles ($K(z) = 1$), and low for shapes that are highly irregular in shape. $K(z)$ depends only on the shape of the zone but is independent of its size. We only report K for the circular scan and SVSS, as this measure evaluates compactness of shapes and cannot be computed over sets of points returned by FSS and ULS. We also introduce an alternate measure of compactness, K_{point}, which operates on sets of points and allows us to compare compactness across all six detection methods. To compute K_{point}, we first find the Voronoi polygons for all spatial locations in the data set, then clip these polygons to the convex hull of the pattern under evaluation and dissolve any shared edges between polygons belonging to points in the pattern. K_{point} is then computed as the area of the polygons covering our pattern divided by the area of the circle with the same total perimeter as these polygons, giving a point-based measure analogous to K. K_{point} is close to 1 for patterns that are roughly circular in shape and not dispersed among points excluded from the pattern. Patterns which are elongated or spread out among excluded points have a low compactness as measured by K_{point}.

Pattern characteristics of the top patterns returned by all methods for three samples are reported in Table 1. Across samples of varying pattern intensity, SVSS scores highest on compactness metrics K and K_{point}. While methods such as FSS and ULS tend to find patterns with higher *LLR* and q_{MLE}, these methods score poorly on compactness, indicating that the detected patterns are sparse and may be sensitive to observations that are elevated due to random noise. With respect to computation time, SVSS is faster than the circular scan and ULS across all pattern intensities, but is slower than DBSCAN, FSS, and kSVM, indicating that the ability to detect spatially compact patterns comes at the expense of an increase in computation time relative to less-constrained detection methods.

Average precision and recall across repeated simulations for patterns with three different signal intensities are reported in Fig. 4. For patterns with a 25% increase in expected counts relative to unaffected points, we find that both SVSS and

Table 1
Summary statistics of detected patterns for simulated regions across three signal intensities: affected regions have a 10%, 25%, and 50% increase in expected counts relative to unaffected regions. Statistics are shown for individual samples from each intensity with affected region shaped like the letter "A".

	n_S			CPU time (s)		
	10%	25%	50%	10%	25%	50%
Circular scan	302	402	496	17.6	17.9	18.2
ULS	744	320	373	55.6	53.7	52.1
FSS	564	511	392	0.29	0.22	0.23
DBSCAN	78	259	377	0.21	0.24	0.2
kSVM	413	550	614	0.55	0.52	0.45
SVSS	236	356	372	15.8	13.3	10.6
	LLR			q_{MLE}		
	10%	25%	50%	10%	25%	50%
Circular scan	74.8	398.8	1579.8	1.07	1.14	1.26
ULS	433.2	1157.1	4017.1	1.09	1.28	1.50
FSS	616.8	1346.6	4041.1	1.18	1.24	1.49
DBSCAN	252.7	1092.6	3734.6	1.27	1.30	1.48
kSVM	540.0	1301.4	3626.8	1.17	1.22	1.36
SVSS	147.3	1067.9	3925.4	1.11	1.25	1.49
	K			K_{point}		
	10%	25%	50%	10%	25%	50%
Circular scan	1.00	1.00	1.00	0.83	0.86	0.88
ULS	–	–	–	0.01	0.03	0.12
FSS	–	–	–	0.01	0.01	0.08
DBSCAN	–	–	–	0.01	0.01	0.06
kSVM	–	–	–	0.01	0.01	0.01
SVSS	0.45	0.48	0.48	0.17	0.13	0.15

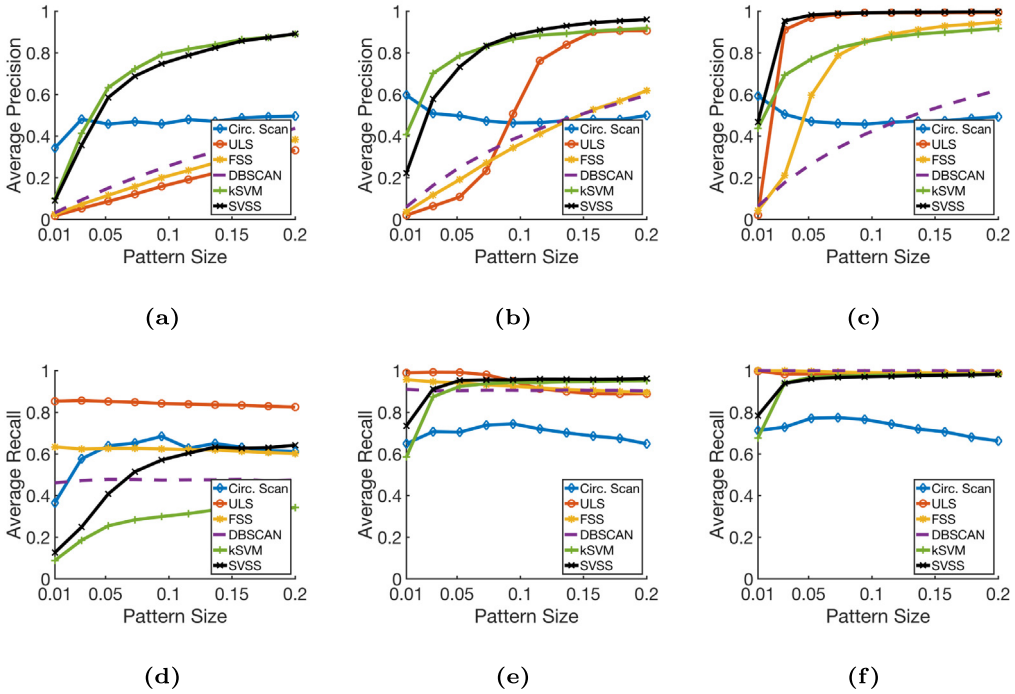

Fig. 4. Average precision (top) and recall (bottom) of six scanning algorithms on detection of letter-shaped patterns of varying size (proportion of study area) in simulated data. Results are shown for three different signal intensities: points in the affected region have a 10% (left), 25% (middle), and 50% (right) increase in expected counts.

kSVM significantly outperform the other methods on precision for the majority of pattern sizes under consideration, indicating that points included in the top SVSS pattern are very likely to be in the true affected region for all but the smallest patterns considered. SVSS outperforms kSVM on pattern sizes larger than 7.5% of the study area. SVSS

Table 2
Summary statistics of top West Nile Virus clusters.

	n_S	CPU time (s)	*LLR*	q_{MLE}	q_{CV}	K	K_{point}
Circular scan	30	0.64	70.8	1.66	1.21	1.00	0.86
ULS	20	0.34	91.8	1.76	–	–	0.10
FSS	25	0.08	116.9	1.84	–	–	0.06
DBSCAN	15	0.16	59.2	1.75	–	–	0.15
kSVM	14	0.16	105.5	1.95	1.32	–	0.10
SVSS	13	7.40	97.3	1.87	1.36	0.62	0.17

demonstrates high recall for patterns large and small, outperforming kSVM and the circular scan across all pattern sizes and outperforming all methods on pattern sizes larger than approximately 10% of the study area. Recall diminishes slightly for the circular scan, FSS, and ULS as patterns increase in size, but SVSS maintains a recall close to 1 even as patterns grow large. Although kSVM demonstrates comparable performance to SVSS with a 25% signal intensity, recall of kSVM drops significantly on patterns with weaker signals.

On weaker patterns with a 10% signal intensity, SVSS and kSVM still outperform competing methods on precision for most of the range of pattern sizes under consideration. Recall of kSVM drops dramatically on the weaker signal relative to signal intensity of 25%. SVSS is beaten by ULS on recall, but significantly outperforms kSVM on recall on the 10% signal intensity. These results suggest that even on relatively weak signals, locations returned by SVSS are very likely to be in the true affected region. The high precision of SVSS on weak signals comes at the expense of reduced recall, but the drop in recall is smaller than for other high-precision methods like kSVM. With stronger signals (e.g., 50% signal intensity), both SVSS and ULS demonstrate high performance across the range of pattern sizes considered and across both evaluation metrics.

3.2. Detecting disease clusters

In the domain of disease surveillance, we demonstrate detection of disease clusters in mosquito pools tested for West Nile Virus (WNV), using data made publicly available by the Chicago Department of Public Health (CDPH) through the City of Chicago Data Portal. Measuring presence of WNV in mosquitoes, a relatively short-lived vector for infection, gives a useful approach to identifying spatial and temporal trends in disease risk throughout a susceptible region (Lampman et al., 2013). Patterns returned by SVSS and other scanning methods indicate the spatial clusters where the proportion of positive test results were elevated with respect to the citywide average over this period, which can help the CDPH target mosquito control programs. Mosquito management is typically implemented through the use of chemical pesticides. Accurately characterizing the spatial regions where the disease is most prevalent in mosquitoes and the risk of transmission to humans is highest can minimize the application of mosquito control measures which may have harmful effects on the ecological health of the treated areas.

Mosquito pools throughout the city are tested regularly for presence of WNV by the CDPH, with individual locations often tested multiple times a year over the course of several years. The expectation-based binomial scan statistic is appropriate in this setting due to the number of total tests varying across spatial locations. Each location thus has an observed count of positive test results c_i, an expected number of positive tests b_i, and a total number of tests n_i. We aggregate observed counts and total number of tests at each test location for a period of over 11 years from June 1, 2007 through September 30, 2018. For the expected number of positive tests, we compute an overall rate of positive test results by aggregating tests across the entire city and the entire study period, then multiply this average rate by the number of total tests n_i at each location. We thus assume a uniform rate of positive test results across test locations under the null hypothesis.

Fig. 5 shows the top patterns detected by six detection algorithms under comparison. The circular scan is constrained in shape and approximates the shape of the true affected region, either with an overly large circle surrounding the affected locations, or with an overly small one identifying only a piece of the affected region. In comparison, SVSS has improved power to detect disease clusters that are elongated or irregular in shape. For example, the top WNV cluster detected by SVSS (Fig. 5f) roughly conforms to sections of two major rivers in North Chicago, overlapping significant portions of the forest preserves adjacent to these rivers. FSS and ULS find patterns that are spread widely throughout the study area and are interspersed with non-anomalous points.

The top patterns returned by each method are characterized in Table 2. SVSS finds a pattern with a higher *LLR* and relative risk q_{MLE} than the circular scan and DBSCAN. For additional validation of the detected patterns, we also compute the *held-out relative risk* q_{CV} by holding out points from the pattern detection methods through 10-fold cross validation, and computing the relative risk of all points that fall within the anomalous pattern decision boundary produced by running the detection method on the other 9 folds. The held-out relative risk values provide evidence that the patterns discovered are meaningful with respect to unseen data or locations not provided to the detection method. For all methods for which we can compute q_{CV} based on the detected decision boundary, we find relatively high values that provide out-of-sample validation of the detected patterns, with SVSS outperforming the circular scan and kSVM on this out-of-sample validation measure.

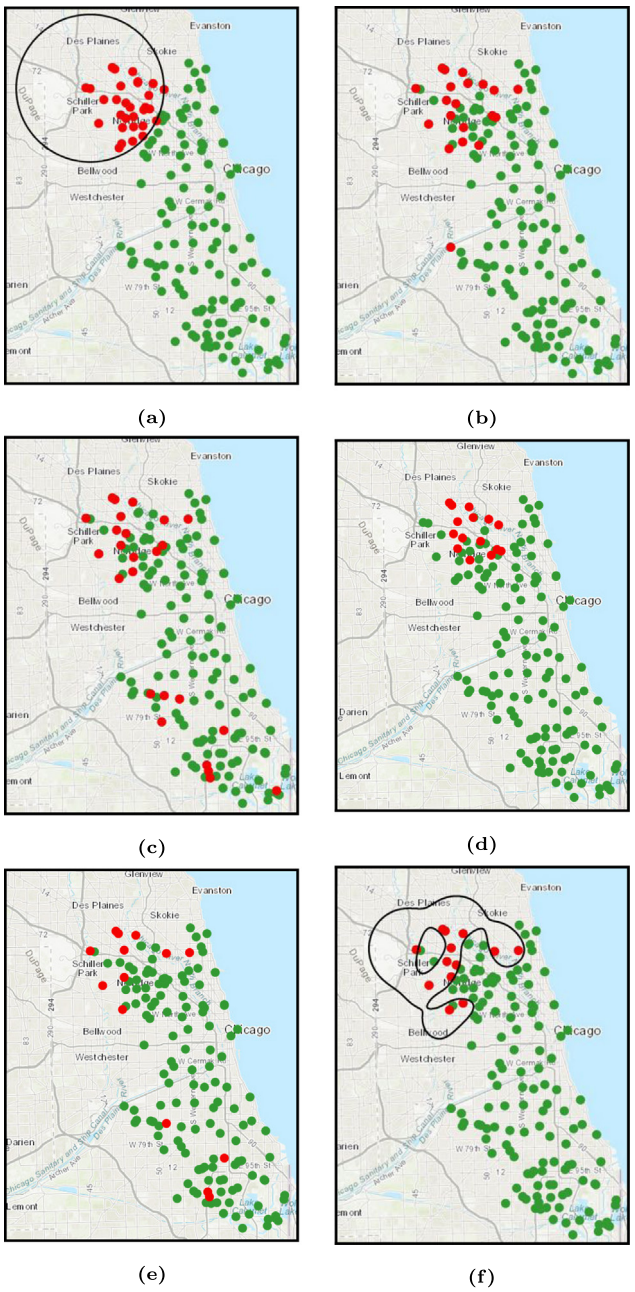

Fig. 5. Clusters of West Nile Virus detected by six pattern detection algorithms in Chicago, IL. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

As measured by K_{point}, the SVSS pattern is more compact than the patterns found by all methods except the circular scan, while still maintaining high relative risk and *LLR* comparable to ULS. FSS finds the unconstrained subset with the highest *LLR* but at the cost of low compactness. While other methods trade off compactness for high *LLR* or vice versa, the pattern returned by SVSS scores highly on both objectives. The higher spatial compactness of the SVSS pattern comes at the cost of higher computation time relative to other methods under comparison.

This analysis applied SVSS in order to detect spatial patterns over a single fixed time window, but the method can be easily extended to track changes in disease hot spots over time by updating the observed and expected values at each location as new data is received. Baseline values can be computed based on pre-outbreak levels, or can be continually updated based on recent trends to assess where hot spots are emerging or spreading. This flexibility in

Table 3
Summary statistics of top street crime clusters.

	n_S	CPU time (s)	LLR	q_{MLE}	Next-year crimes/cell	K	K_{point}
Circular scan	347	36.4	257.4	1.26	29.9	1.00	0.910
ULS	1102	171.7	986.3	1.39	17.0	–	0.005
FSS	945	0.3	1687.5	1.77	11.0	–	0.002
DBSCAN	536	0.2	1311.2	2.69	4.8	–	0.003
kSVM	1252	6.6	1310.3	1.48	13.7	–	0.003
SVSS	115	379.6	420.0	1.62	32.4	0.23	0.027

definition of baseline values makes SVSS well-suited for problem settings where it is necessary to characterize the changes in anomalous patterns over time.

3.3. Detecting crime hot-spots

Next, we apply SVSS in the context of crime surveillance using calls-for-service records from Portland, OR. These records were made publicly available by the Portland Police Bureau (PPB) for the National Institute of Justice's Real-Time Crime Forecasting Challenge. We restrict our analysis to calls-for-service relating to "street crime" as categorized by the PPB, which includes assaults, robberies, shootings, stabbings, and vice-related crimes, among other crime types. We aggregate geotagged CFS records to 1000-foot square grid cells, and estimate location-specific expected counts using the time series for each cell. Specifically, we compute expected counts as an average annual count of street crimes for each grid cell using data from the three year period from March 2012 through February 2015. Observed counts are aggregated over the following year, from March 1, 2015 through February 29, 2016. We use the expectation-based Poisson scan statistic for all six methods. Patterns returned by SVSS and other scanning methods indicate spatial regions where observed crime in the most recent year of data was elevated relative to expected counts estimated from the previous three years. Such regions could indicate newly emerging hot-spots of crime, e.g., due to changing neighborhood composition, new patterns of gang or other criminal activity, crime attractors such as bars or liquor stores, or other structural changes. While police departments are typically aware of neighborhoods with chronically high levels of crime, they may not be aware of newly emerging hot-spots which could be effectively targeted for crime prevention.

The crime patterns from six pattern detection algorithms are displayed in Fig. 6, with summary statistics reported in Table 3. The circular scan finds a circular pattern covering much of Downtown and East Portland on either side of the Willamette River. The SVSS pattern is situated in roughly the same area of Southeast Portland as the circular scan pattern, but is highly irregular in shape and extends eastward to encompass the Hawthorne District, a popular commercial strip known for its bohemian vibe and vintage clothing stores. While not a particularly high-crime area as compared to downtown Portland, the high foot traffic and store density in this area provide ample opportunity for larceny that could be prevented through targeted police patrols. The SVSS pattern has higher *LLR* and relative risk when compared with the circular scan. FSS and ULS both result in large patterns that span most of the city, with higher *LLR* than SVSS but extremely low relative compactness. The large size and relative sparsity of these patterns indicate that FSS and ULS are badly overfitting. The methods are not sufficiently constrained to produce coherent subsets, so they pick out many isolated points throughout the study region with high counts due to chance. As an additional evaluation metric, we report the count of street crimes per cell for the year following the test period, from March 1, 2016 through February 28, 2017. We find that the SVSS pattern resulted in the highest crimes per cell across all six methods in the year following the test period. Even though FSS and ULS pick out points with high relative risk in the training data (comparable to SVSS), the points chosen by SVSS have much higher crime rate in the subsequent year's data and thus seem to be a much better target for proactive police patrols.

3.4. Detecting pothole clusters

For our final application, we apply SVSS in the domain of city services and management to detect clusters of pothole complaints in Pittsburgh, PA. Our data set for this analysis consists of publicly available call records from Pittsburgh's 311 system. People living in Pittsburgh can call the 311 telephone center to notify the city of any non-emergency issues, including requests for service related to road deterioration. Potholes represent one of the most common issues reported to the city, with pothole reports making up 13.1% of all 311 calls between 2016 and 2018. Detecting clusters in these reports has the potential to help public works agencies in Pittsburgh and other cities identify and efficiently respond to emerging clusters of potholes.

We aggregate counts of pothole reports to city blocks, using a two-year period from January 1, 2016 through December 31, 2017 to estimate a city-wide average annual count of potholes. Similar to the disease outbreak detection application, we thus assume a uniform baseline rate of pothole reports under the null hypothesis. We find observed counts for each city block from January 1, 2018 through December 31, 2018, and apply the expectation-based Poisson scan statistic to search for spatial regions with elevated counts of potholes in 2018 in comparison to the previous two years. Such clusters could indicate newly emerging regions in need of attention due to weather events or recent shifts in traffic patterns

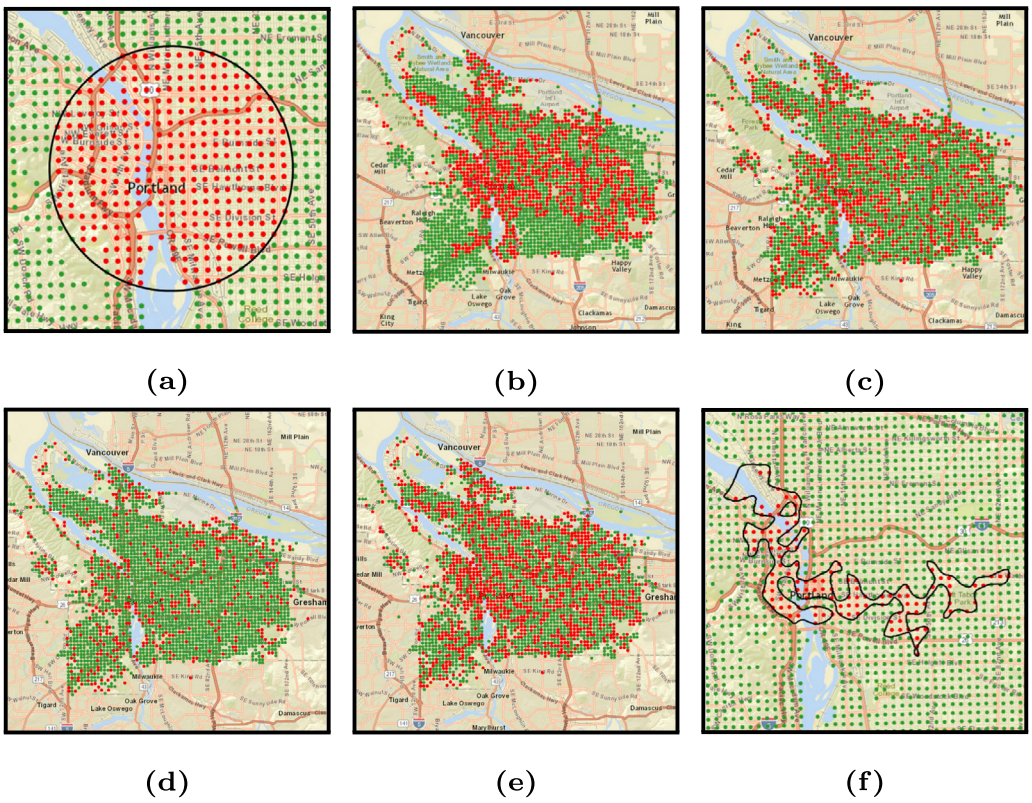

Fig. 6. Clusters of street crime detected by six pattern detection algorithms in Portland, OR. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

Table 4
Summary statistics of top pothole clusters.

	n_S	CPU time (s)	*LLR*	q_{MLE}	q_{CV}	K	K_{point}
Circular scan	497	28.3	2038.0	4.48	4.15	1.00	0.876
ULS	1096	140.3	6128.3	5.22	–	–	0.006
FSS	642	0.3	9182.2	8.78	–	–	0.003
DBSCAN	1607	0.2	7635.4	4.81	–	–	0.003
kSVM	1805	2.0	4242.2	3.48	3.26	–	0.076
SVSS	111	131.4	2272.3	10.91	4.12	0.42[a]	0.030[a]

[a]Denotes average over top five disjoint components.

contributing to road surface deterioration, helping public works agencies plan and prioritize future road maintenance projects.

For this analysis, we demonstrate an alternate approach to finding irregular patterns with SVSS that may have multiple disconnected regions. In many real-world use cases for pattern detection, multiple affected regions exist in the same data and we therefore would benefit from a method for both detecting and prioritizing over many anomalous clusters. If desired for operational purposes, SVSS allows users to rank the disconnected regions by the unpenalized log-likelihood ratio statistic and choose k components to include in order to retrieve an anomalous pattern of the desired scale (Higher k produces a larger pattern consisting of more disconnected but individually compact regions). Instead of selecting the single top component from the connected components of the SVM decision boundary (as discussed in Section 2.4), here we include the top 5 disconnected components of the pattern returned by SVSS. Public works agencies could scale a proposed infrastructure project up or down based on operational constraints by increasing or decreasing the number of disjoint components to include.

Fig. 7 displays the top pothole clusters returned by six pattern detection methods, and Table 4 provides summary statistics for these patterns. For the compactness measures, we report the average compactness across the top 5 components for SVSS. As discussed above, SVSS returns a pattern consisting of multiple disconnected regions. This pattern has the highest relative risk q_{MLE} among the detection methods under comparison, and higher *LLR* than the circular scan pattern. Although SVSS has very high q_{MLE} relative to other methods, q_{CV} reveals a held-out relative risk that is more in

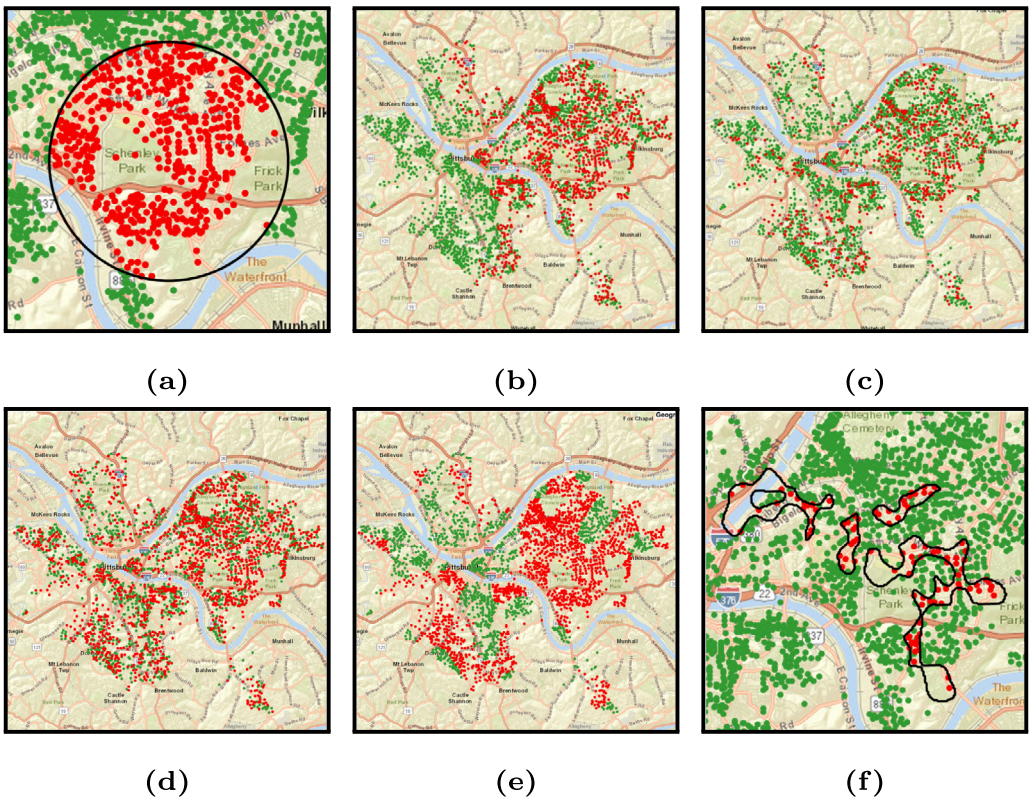

Fig. 7. Clusters of potholes detected by six pattern detection algorithms in Pittsburgh, PA. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

line with competitors, possibly indicating that q_{MLE} is heavily influenced by high counts at a small number of individual points in the full data set. Still, the held-out relative risk q_{CV} of both SVSS and the circular scan are comparable and higher than that of kSVM. The individual components of SVSS correspond to highly trafficked roads and intersections throughout Pittsburgh that are subject to high rates of wear and degradation, with 4 of the 5 components overlapping one or more public bus routes. The disconnected regions which make up the SVSS pattern are elongated due to the underlying spatial structure of the road network. Yet these regions are still individually compact, as indicated by the high average geometric compactness measures relative to the sparse and underconstrained patterns found by ULS and FSS. As in the previous two applications, SVSS scores relatively highly on *both* compactness and measures of anomalousness, resulting in patterns that are highly anomalous but still spatially coherent.

3.5. Discussion of real-world case studies

In all three of the above case studies, the literature reveals multiple distinct environmental factors that can drive West Nile Virus, crime, or potholes respectively. Thus, these factors do not clearly indicate which part of the city to target with public health, law enforcement, or road maintenance interventions respectively, while our approach precisely localizes a spatial area that can benefit from targeted intervention.

For West Nile Virus, *Culex* species mosquitoes which transmit the virus can breed in a variety of stagnant water sources, including low places with poor drainage, urban catch basins, roadside ditches, sewage treatment lagoons, and manmade containers around houses (Ruiz et al., 2007). A variety of other factors including temperature, humidity, rainfall, surface permeability, and bird migration patterns have been identified as predictive (Hernandez et al., 2019). Human WNV cases in a 2002 outbreak in Chicago were found to be associated with higher percentages of vegetation in a census tract, and areas in Chicago's inner suburbs were found to have higher human WNV rates than either the outer suburbs or the urban center (Ruiz et al., 2007). Thus the prior literature supports our identification of certain forest preserve and river areas as WNV hot spots but does not necessarily point to these particular areas in North Chicago. Similarly, the literature on crime prediction reveals that chronic hot spots of crime are often found in large commercial areas and nearby residential areas (Fitzpatrick et al., 2019), and while the Hawthorne District is one well-known commercial district of Portland, there was no reason to expect *a priori* that this particular strip would exhibit a flare-up of property crimes in the particular

year of data under analysis. Finally, predictive factors for pothole formation include weather (temperature and freeze-thaw cycles), pavement condition, and traffic loads (Sadeghi et al., 2016). An analysis by the Metropolitan Transportation Commission (The Pothole Report, 2011) estimates that buses and other large vehicles create thousands of times more physical stress on pavements per trip as compared to passenger vehicles, supporting our discovery of spatial clusters of potholes on heavily trafficked bus routes in Pittsburgh.

4. Conclusions

In this paper, we introduce the support vector subset scan (SVSS), a novel method for detecting anomalous patterns in spatial data that are spatially compact and irregular in shape. SVSS integrates soft spatial constraints into the fast subset scan, rewarding patterns with spatial coherence. As demonstrated above in the contexts of disease outbreak detection, crime surveillance, and city services and management, SVSS provides a flexible framework for spatial pattern detection in a variety of problem settings where detection and characterization of coherent anomalous patterns in spatial data has demonstrable real-world benefits.

SVSS enables discovery of compact, anomalous patterns in spatial data, but these desirable properties come at the expense of increased computation time relative to other methods, due to the iteration between subset scanning and SVM learning steps. As a result of this limitation, SVSS is best suited for applications where a moderate increase in computation time is acceptable in order to properly characterize the boundaries of an anomalous region with high fidelity and spatial resolution. SVSS is also most appropriate and effective for cases where the true clusters of interest are likely to be irregular in shape yet spatially compact. In the extreme case where clusters tend to be nearly circular in shape, we expect SVSS and other flexible scan methods to underperform the circular scan with respect to detection power and spatial accuracy, due to the larger search space. On the other extreme, where the patterns of interest are highly dispersed and do not cluster spatially, we expect SVSS and other constrained scan methods to underperform the unconstrained scan.

An interesting avenue for further research is the development of combined evaluation metrics for irregularly-shaped cluster detection that incorporate both anomalousness and spatial compactness. While the relative weighting of these two components might be highly domain-specific, we could optimize a linear combination of an anomalousness measure and a compactness measure (for example, a penalized log-likelihood ratio). Alternatively, we could optimize the anomalousness measure subject to a hard lower-bound constraint on compactness, and optionally, hard constraints on the number and/or size of clusters. In either case, we recommend focusing on out-of-sample performance with respect to anomalousness (e.g., via cross-validation or assessment on a separate held-out time period), so that methods are not rewarded for overfitting. We also recommend using measures that can be computed across the range of methods being considered (e.g., using K_{point} instead of K to evaluate compactness). Characteristics of patterns returned by SVSS may also be helpful as features in predictive models related to the spatial data in question. For example, grid cells returned by SVSS as part of crime clusters reported more crime in the following year than those returned by other pattern detection methods. In future work, the authors plan to further evaluate how inclusion of SVSS cluster attributes can improve prediction models in areas of public health and safety.

Acknowledgments

This work was partially funded by NSF, USA grant IIS-0953330. A preliminary version was presented at the International Society for Disease Surveillance Annual Conference with a one-page abstract published in the *Online Journal of Public Health Informatics* (Fitzpatrick et al., 2017).

Appendix A. Implementation details

As discussed in Section 2.1, Speakman et al. (2016) provide the expressions for the log-likelihood ratio statistics $\lambda_i(q)$ for the expectation-based binomial scan statistic (EBB, used in Section 3.2), the expectation-based Poisson scan statistic (EBP, used in Sections 3.1, 3.3, and Section 3.4) and others in the exponential family. We include the expressions for $\lambda_i(q)$ in Table 5 for ease of reproducibility. We also report optimized parameter values for the three SVSS tuning parameters used in the experiments in Sections 3.2–3.4 in Table 6.

To evaluate our method in an experimental setting, we generate patterns of varying size, shape, and intensity in simulated data, and compare precision and recall of SVSS with five other methods for spatial pattern detection:

- The *circular scan statistic*, which searches over N^2 total circles and returns the circle with the highest log-likelihood ratio (LLR). For each location, we evaluate the N circles of increasing radius centered at the location, such that each successive circle grows to include one additional neighboring point (Kulldorff, 1997).
- The *upper level set scan statistic (ULS)*, which searches over connected components of all possible upper level sets with respect to the ratio of observed values to baselines. ULS searches over tessellated cells rather than points, so we construct a Voronoi tessellation from points in space as a pre-processing step (Patil and Taillie, 2004).
- The *fast subset scan (FSS)*, which returns the subset of locations which maximizes the unpenalized LLR (Neill, 2012).

Table 5
Location-specific contributions to the score function for expectation-based statistics in the exponential family. See Speakman et al. (2016) for full derivations.

Distribution	$\lambda_i(q)$
Poisson	$c_i \log q + b_i(1-q)$
Gaussian	$c_i b_i \frac{(q-1)}{\sigma_i^2} + b_i^2(\frac{1-q^2}{2\sigma_i^2})$
Exponential	$\frac{c_i}{b_i}(1-\frac{1}{q}) - \log q$
Binomial	$c_i \log q + (n_i - c_i)\log(\frac{n_i - qb_i}{n_i - b_i})$
Negative binomial	$c_i \log q + (r_i + c_i)\log(\frac{r_i + b_i}{r_i + qb_i})$

Table 6
Parameter values for application of SVSS on real world data sets.

Data set	Gaussian kernel bandwidth	C_0	C_1
Chicago West Nile	0.09	50	100
Portland street crime	0.03	100 000	200 000
Pittsburgh potholes	0.03	1000	2000

- *DBSCAN with thresholding*, a clustering algorithm that finds high-density clusters of arbitrary shape (Ester et al., 1996). Only with locations with count-to-baseline ratios above a fixed threshold are clustered. The threshold and DBSCAN parameters are selected to optimize anomalousness (LLR) of the top cluster. The single cluster with highest LLR is considered as the pattern returned by DBSCAN.
- *Kernel Support Vector Machine (kSVM) with thresholding*, which applies a threshold to the ratio of counts to baselines for each location, then trains an SVM with a Gaussian kernel to learn a nonlinear decision boundary between points above and below the threshold. The threshold and SVM parameters are chosen using 10-fold cross-validation to optimize the anomalousness score (LLR).

Appendix B. Supplementary data

For ease of reproducibility, the three real-world data sets described in Sections 3.2–3.4 are all publicly available and any processed versions of the data sets are posted as supplementary material alongside our code.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.107149.

References

Bottou, L., Lin, C.-J., 2007. Large-Scale Kernel Machines: Support Vector Machine Solvers. MIT Press.
Cancado, A.L., Duarte, A.R., Duczmal, L.H., Ferreira, S.J., Fonseca, C.M., Gontijo, E.C., 2010. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. Int. J. Health Geogr. 9 (55).
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Costa, M.A., Kulldorff, M., 2014. Maximum linkage space–time permutation scan statistics for disease outbreak detection. Int. J. Health Geogr. 13 (20).
Diggle, P., Tawn, J., Moyeed, R., 2002. Modelbased geostatistics. J. R. Stat. Soc. (Ser. C: Appl. Stat.) 47 (3).
Duarte, A.R., Duczmal, L., Ferreira, S.J., 2010. Internal cohesion and geometric shape of spatial clusters. Environ. Ecol. Stat. 17 (2), 203–229.
Duczmal, L., Assuncao, R., 2004. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. Comput. Statist. Data Anal. 45, 269–286.
Duczmal, L.H., Cançado, A., Takahashi, R.H.C., 2012. Delineation of irregularly shaped disease clusters through multiobjective optimization. J. Comput. Graph. Statist. 2008 (1), 243–262.
Duczmal, L., Cancado, A.L., Takahashi, R.H., Bessegato, L., 2007. A genetic algorithm for irregularly shaped spatial scan statistics. Comput. Statist. Data Anal. 52, 43–52.
Duczmal, L., Kulldorff, M., Huang, L., 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. J. Comput. Graph. Statist. 15 (2), 428–442.
Ester, M., Kriegal, H., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231.
Fitzpatrick, D.J., Gorr, W.L., Neill, D.B., 2019. Keeping score: predictive analytics in policing. Annu. Rev. Criminol. 2, 473–491.
Fitzpatrick, D.J., Ni, Y., Neill, D.B., 2017. Support vector subset scan for spatial outbreak detection. Online J. Public Health Inform. 9 (1), e021.
Gorski, J., Pfeuffer, F., Klamroth, K., 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. Math. Methods Oper. Res. 66 (3), 373–407.
Haran, M., 2011. Gaussian random field models for spatial data. Handb. Markov Chain Monte Carlo 449–478.
Hernandez, E., Torres, R., Joyce, A., 2019. Environmental and sociological factors associated with the incidence of west nile virus cases in the northern san joaquin valley of california, 2011–2015. Vector-Borne Zoonotic Dis. 19 (11), 851–858.
Kulldorff, M., 1997. A spatial scan statistic. Comm. Statist. Theory Methods 26 (6), 1481–1496.
Kulldorff, M., Huang, L., Pickle, L., Duczmal, L., 2006. An elliptical spatial scan statistic. Stat. Med. 25, 3929–3943.
Lampman, R.L., Krasavin, N.M., Ward, M.P., Beveroth, T.A., Lankau, E.W., Alto, B.W., Muturi, E., Novak, R.J., 2013. West nile virus infection rates and avian serology in east-central illinois. J. Amer. Mosq. Control Assoc. 29 (2), 108–122.

Moreira, G.J.P., Paquete, L., Duczmal, L.H., Menotti, D., Takahashi, R.H.C., 2015. Multi-objective dynamic programming for spatial cluster detection. Environ. Ecol. Stat. 22 (2), 369–391.
Neill, D.B., 2009. Expectation-based scan statistics for monitoring spatial time series data. Int. J. Forecast. 25, 498–517.
Neill, D.B., 2012. Fast subset scan for spatial pattern detection. J. R. Stat. Soc. (Ser. B: Stat. Methodol.) 74 (2), 337–360.
Neill, D.B., Moore, A.W., 2004. Rapid detection of significant spatial clusters. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 256–265.
Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K., 2005. Detection of emerging space–time clusters. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 218–227.
Patil, G., Taillie, C., 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ. Ecol. Stat. 11, 183–197.
Ruiz, M., Walker, E., Foster, E., Haramis, L., Kitron, U., 2007. Association of west nile virus illness and urban landscapes in chicago and detroit. Int. J. Health Geogr. 6 (10).
Sadeghi, L., Zhang, Y., Balmos, A., Krogmeier, J., Haddock, J., 2016. Algorithm and Software for Proactive Pothole Repair (Joint Transportation Research Program Publication No. Fhwa/in/Jtrp-2016/14). Tech. rep., Purdue University, West Lafayette, IN.
Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13 (7), 1443–1471.
Speakman, S., McFowland, E., Neill, D.B., 2015. Scalable detection of anomalous patterns with connectivity constraints. J. Comput. Graph. Statist. 24 (4), 1014–1033.
Speakman, S., Somanchi, S., McFowland, E., Neill, D.B., 2016. Penalized fast subset scanning.. J. Comput. Graph. Statist. 25 (2), 382–404.
Takahashi, K., Kulldorff, M., Tango, T., Yih, K., 2008. A flexibly shaped space–time scan statistic for disease outbreak detection and monitoring. Int. J. Health Geogr. 7 (14).
2011. The Pothole Report: Can the Bay Area Have Better Roads?. Tech. rep., Metropolitan Transportation Commission.
Yiannakoulias, N., Rosychuk, R.J., Hodgson, J., 2007. Adaptations for finding irregularly shaped disease clusters. Int. J. Health Geogr. 6 (28).