group on knowledge discovery and data mining, Boston, pp 71–80
Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. MIT, Cambridge
Hershberger J, Shrivastava N, Suri S (2006) Cluster hulls: a technique for summarizing spatial data streams. In: Proceedings of IEEE international conference on data engineering, Atlanta, p 138
Hulten G, Spencer L, Domingos P (2001) Mining time-changing data streams. In: Proceedings of ACM special interest group on knowledge discovery and data mining, San Francisco, pp 97–106
Natwichai J, Li X (2004) Knowledge maintenance on data streams with concept drifting: international symposium on computation and information sciences (CIS), Shanghai, pp 705–710
O'Callaghan L, Mishra N, Meyerson A, Guha S, Motwani R (2002) Streaming-data algorithms for high-quality clustering. In: Proceedings of IEEE international conference on data engineering, San Jose, pp 685–694
Pan F, Wang B, Ren D, Hu X, Perrizo W (2003) Proximal support vector machine for spatial data using peano trees. In: ISCA computer applications in industry and engineering, Las Vegas, pp 292–297
Perrizo W, Jockheck W, Perera A, Ren D, Wu W, Zhang Y (2002) Multimedia data mining using P-trees. In: International workshop on multimedia data mining (MDM/KDD), Edmonton, pp 19–29
Rao KR, Yip P (1990) Discrete cosine transform: algorithms, advantages, applications. Academic, San Diego
Ruoming J, Agrawal G (2003) Efficient decision tree construction on streaming data. In: ACM special interest group on knowledge discovery and data mining (SIGKDD), Washington, DC, pp 571–576
Versavel J (1999) Road safety through video detection. In: Proceedings of IEEE international conference on intelligent transportation system, Boulder, pp 753–757
Wang B, Pan F, Ren D, Cui Y, Ding Q, Perrizo W (2003) Efficient OLAP operations for spatial data using peano trees: In ACM special interest group on management of data workshop (SIGMOD), San Diego, pp 28–34
Yao Y, Gehrke JE (2002) The cougar approach to in-network query processing in sensor networks: ACM special interest group on data management of data (SIGMOD) record, pp 9–18
Zhao J, Lu CT, Kou Y (2003) Detecting region outliers in meteorological data. In: Proceedings of ACM international symposium on advances in geographic information system, New Orleans, pp 49–55

## StreamInsight

▸ Data Stream Systems, Empowering with Spatiotemporal Capabilities

## Streams

▸ Queries in Spatiotemporal Databases, Time Parameterized

## Subset Scanning for Event and Pattern Detection

Daniel B. Neill
Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA

### Synonyms

Fast subset scan; GraphScan; Linear-time subset scanning

### Definition

Subset scanning is an accurate and computationally efficient framework for detecting events and other patterns in both spatial and nonspatial datasets, through constrained optimization of a score function (e.g., a likelihood ratio statistic) over subsets of the data. Many score functions of interest satisfy the *linear-time subset scanning* property (Neill 2012), enabling exact and efficient optimization over subsets. This efficient unconstrained optimization step, the *fast subset scan*, can be used as a building block for scalable solutions to event and pattern detection problems incorporating a variety of real-world constraints.

### Historical Background

The spatial and space-time scan statistics (Kulldorff 1997, 2001), building on earlier work on scan statistics by Naus (1965) and others, are powerful and widely used methods for event detection in spatiotemporal data. These methods evaluate a score function $F(S)$, typically a likelihood ratio statistic, over a large set of spatial or space-time regions $S$, identifying

those regions which are most likely to represent anomalous spatial clusters. A typical approach is to search over all regions of a given shape such as circles (Kulldorff 1997), space-time cylinders (Kulldorff 2001), rectangles (Neill and Moore 2004), or ellipses (Kulldorff et al. 2006). These approaches perform well when the true spatial region ("cluster") of interest is well approximated by the set of search regions but suffer from reduced detection power otherwise. For example, the original spatial scan approach (Kulldorff 1997), searching over circular regions, loses power for highly elongated clusters, and all of the above approaches lose power for clusters with highly irregular shapes. Many recent approaches search over larger sets of irregularly shaped regions, such as subsets of locations connected by spatial adjacency (Patil and Taillie 2004; Tango and Takahashi 2005; Duczmal and Assuncao 2004). Most of these approaches perform approximate rather than exhaustive searches over the chosen set of irregularly shaped regions, using some heuristic optimization approach such as simulated annealing (Duczmal and Assuncao 2004) or genetic algorithms (Duczmal et al. 2007), and thus do not guarantee that an optimal or even near-optimal region will be found. Alternative approaches search exhaustively over a much smaller set of regions based on upper level sets (Patil and Taillie 2004) or spanning trees (Costa et al. 2012), but again, these approaches may fail to identify the highest-scoring subset. Finally, the FlexScan approach of Tango and Takahashi (2005) performs an exhaustive search over connected subsets within the local neighborhood of each spatial location but is computationally expensive and does not scale to even moderately large neighborhood sizes. These limitations of previous methods led to the development of the fast subset scan (Neill 2012), enabling exact and efficient search over subsets of locations. Recent extensions of fast subset scanning allow incorporation of various constraints including spatial proximity and connectivity. Moreover, empirical comparisons suggest that fast subset scanning approaches outperform competing methods with respect to both computational efficiency (scaling to much larger datasets) and detection power.

## Scientific Fundamentals

### Additional Motivation for the Subset Scanning Approach

As noted above, subset scanning originated from the spatial statistics literature, building on Kulldorff's spatial scan approach (Kulldorff 1997) in order to accurately detect irregularly shaped spatial clusters or subsets satisfying other relevant constraints (e.g., anomalous subgraphs of a larger real-world network). It has since been generalized beyond the spatial domain to identify subsets of similar data records for which some subset of attributes are anomalous (McFowland III et al. 2013), as well as to many other domains, e.g., detecting patterns in massive image data (Somanchi and Neill 2013) and detecting events using online social network data such as Twitter (Chen and Neill 2014).

From a machine learning and data mining perspective, the idea of detecting subsets that are collectively interesting or anomalous is a natural extension of typical, single record-based anomaly detection approaches (Das et al. 2008). However, most previous approaches to this problem are either heuristic search methods (which are not guaranteed to find optimal or approximately optimal subsets), top-down detection methods (which search for globally interesting patterns and then drill down to more carefully investigate the most interesting sub-partitions), or bottom-up detection approaches (which find individually anomalous data points and then aggregate them into clusters). However, top-down methods often fail to detect *small-scale patterns* that may not be evident from the global aggregate statistics, while bottom-up methods can fail for *subtle patterns* that are only evident when a sufficiently large group of data records are considered collectively. For example, a subtle increase in the number of emergency department visits in several nearby hospitals may be indicative of an emerging disease outbreak, but this signal might not be visible when observing only a single hospital or when

aggregating visit counts across all hospitals in a given area.

Of course, subset scanning creates both statistical and computational challenges, the most serious of which is the computational infeasibility of exhaustively searching over the exponentially many subsets. This computational challenge has been addressed by the fast subset scan approach described below, which exploits the "linear-time subset scanning" property of many commonly used score functions to perform exact and efficient search over subsets. A second, statistical challenge is that multiple testing may result in large numbers of false-positives, particularly when searching over a huge number of subsets. Randomization testing can be used within the spatial and subset scanning framework to bound the overall number of false-positives under the null hypothesis $H_0$ but is computationally expensive and can still result in high false-positive rates when $H_0$ is mis-specified. An alternative approach that mitigates these problems is empirical calibration using historical background data (Neill 2009a; Chen and Neill 2014). In either case, an *under-constrained search* over an excessive number of subsets can lead to higher threshold scores required for detection at a given false-positive rate, reducing detection power, while an *over-constrained search* loses detection power whenever the true affected region falls outside the search space. This motivates the recent development of constrained fast subset scan approaches that can incorporate relevant real-world constraints such as spatial proximity and graph connectivity, leading to substantially improved detection power.

### Likelihood Ratio Statistics and the Spatial Scan

In a typical formulation of the multivariate event detection problem, we are given a dataset $D$ consisting of multiple data streams $D_m$ ($m = 1 \ldots M$) monitored at a set of spatial locations $s_i$ ($i = 1 \ldots N$). For each combination of stream $D_m$ and location $s_i$, we are given a time series of observed counts $c^t_{i,m}$. In the *expectation-based scan statistic* framework (Neill et al. 2005; Neill 2009b), we also infer an expected count $b^t_{i,m}$ corresponding to each observed count $c^t_{i,m}$, typically by time series analysis of historical data. Given all of this data, a typical goal is to identify spatial regions where the recent counts for some subset of the monitored data streams are significantly higher than expected. For example, in disease surveillance, we monitor a variety of health-related data streams, such as emergency department visits (for different symptom categories) and over-the-counter medication sales (for different product categories) and search for spatial areas with emerging overdensities of disease.

We focus first on the univariate case, in which we have only a single monitored data stream ($M = 1$). The spatial and space-time scan statistics are commonly used methodological approaches to this problem, where we define a set of search regions $S$ and evaluate a score function $F(S)$ for each region. The highest-scoring regions are considered to be the most likely clusters, and statistical significance of each region can be determined by randomization testing; see Kulldorff (1997) for details. In the original spatial scan approach (Kulldorff 1997), the set of search regions $S$ is assumed to be the $N^2$ distinct circular regions centered at each of the $N$ locations and consisting of that location and its $k - 1$ nearest neighbors, for $k = 1 \ldots N$. For the space-time scan (Kulldorff 2001), the time duration $W$ of each cluster is also allowed to vary between 1 and some maximum temporal window size $W_{\max}$, resulting a larger set of $N^2 W_{\max}$ cylindrical space-time regions. The score function is typically a *log-likelihood ratio statistic* that incorporates parametric models of how counts are generated both under the null hypothesis $H_0$, assuming no clusters, and the alternative hypothesis $H_1(S)$, assuming a cluster in region $S$. Given these models, the log-likelihood ratio score is defined as:

$$F(S) = \log\left(\frac{\Pr(D \mid H_1(S))}{\Pr(D \mid H_0)}\right).$$

For the expectation-based scan statistics (Neill et al. 2005; Neill 2009b), the null hypothesis $H_0$ assumes that each count $c^t_i$ is drawn from some parametric distribution in a single-parameter

exponential family, with mean equal to $b_i^t$. The alternative hypothesis $H_1(S)$ assumes a constant multiplicative increase $q > 1$ for all expected counts in the region $S$. For example, for the expectation-based Poisson (EBP) scan statistic (Neill 2009b), we have $c_i^t \sim \text{Poisson}(b_i^t)$ everywhere under $H_0$. For $H_1(S)$, we have $c_i^t \sim \text{Poisson}(qb_i^t)$ for all counts inside region $S$ and $c_i^t \sim \text{Poisson}(b_i^t)$ outside $S$. The maximum likelihood estimate of $q$ is $\max\left(1, \frac{C(S)}{B(S)}\right)$, where $C(S) = \sum_S c_i^t$ and $B(S) = \sum_S b_i^t$. Plugging this value of $q$ and the Poisson likelihoods into the equation above results in the EBP score function:

$$F_{\text{EBP}}(S) = C(S) \log\left(\frac{C(S)}{B(S)}\right) + B(S) - C(S),$$

if $C(S) > B(S)$ and $F(S) = 0$ otherwise. The EBP score function is slightly different than Kulldorff's original Poisson scan statistic (Kulldorff 1997) and has the advantage of high detection power for both large and small clusters, while Kulldorff's statistic loses detection power for large clusters (Neill 2009a). More generally, for expectation-based scan statistics in a *separable exponential family*, including the Poisson, Gaussian, and exponential distributions, the score function can be written in the form $F(S) = B(S)D_\phi\left(\frac{C(S)}{B(S)}, 1\right)$, where $D_\phi$ is a Bregman divergence. See Neill (2012) for more details.

### Linear-Time Subset Scanning and the Fast Subset Scan

As noted above, typical spatial and space-time scan approaches suffer from reduced detection power when the true affected subset of locations does not correspond well to the set of search regions, e.g., for elongated or irregularly shaped clusters. Detection power can be substantially improved by optimizing the score function $F(S)$ over all subsets $S$, typically with additional constraints (such as spatial proximity) to ensure that the discovered clusters are feasible solutions to the problem under consideration. However, an exhaustive search over subsets, evaluating the score function $F(S)$ for all $2^N$ subsets $S \subseteq \{s_1 \ldots s_N\}$, quickly becomes computationally infeasible. This motivated the development of fast subset scanning approaches designed to find the highest-scoring subsets $S^* = \arg\max_S F(S)$ without an exhaustive search.

The key idea underlying these approaches, as described in Neill (2012), is that many relevant score functions satisfy a property (linear-time subset scanning or LTSS) that allows efficient optimization over subsets, by sorting the spatial locations $s_i$ according to some "priority" function $G(s_i)$ and evaluating only those subsets consisting of the top-$j$ highest priority locations, for $j = 1 \ldots N$. For functions satisfying the LTSS property, $\max_S F(S) = \max_j F(\{s_{(1)} \ldots s_{(j)}\})$, where $s_{(j)}$ is the $j$th highest priority location, and thus, we are guaranteed that the highest scoring of all $2^N$ subsets will be one of the $N$ subsets that are evaluated. This *fast subset scan* approach dramatically reduces computation time while still guaranteeing an exact solution to the unconstrained (all subsets) search problem. For example, the highest-scoring subset of 97 zip codes in Allegheny County, Pennsylvania, can be found in approximately 40 ms, while an exhaustive search would require about $10^{20}$ years (Neill 2012). The fact that an exact, rather than approximate, solution is found makes fast subset scanning fundamentally different from other approaches based on submodular function optimization (Leskovec et al. 2007), which produce provably good approximations but do not necessarily identify the optimal subset.

The linear-time subset scanning property has been shown to hold for many useful score functions, including both parametric log-likelihood ratio statistics (such as the expectation-based scan statistics and Kulldorff's original spatial scan statistic) and nonparametric scan statistics (McFowland III et al. 2013). Following Neill (2012), we consider three different conditions, each of which is sufficient for the LTSS property to hold:

- Let $F(S) = F(X(S), Y(S))$ be a convex (or quasi-convex) function of two additive,

nonnegative sufficient statistics of subset $S$, $X(S) = \sum_{s_i \in S} x_i$ and $Y(S) = \sum_{s_i \in S} y_i$. If $F(S)$ is monotonically increasing with $X(S)$ or decreasing with $Y(S)$, then $F(S)$ satisfies the LTSS property with priority function $G(s_i) = \frac{x_i}{y_i}$. The key step in this proof is to show that, if there exist two locations $s_{\text{in}} \in S$ and $s_{\text{out}} \notin S$, where $G(s_{\text{in}}) \leq G(s_{\text{out}})$, then $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$, i.e., the score of subset $S$ will be increased by either adding the higher-priority location $s_{\text{out}}$ or removing the lower-priority location $s_{\text{in}}$. This step follows from the convexity of function $F$. As a corollary, the original formulation of the spatial scan statistic (Kulldorff 1997) satisfies LTSS.

- Let $F(S) = F(X(S), |S|)$ be a function of one additive sufficient statistic of subset $S$ and the cardinality of $S$. If $F(S)$ is monotonically increasing with $X(S)$, then $F(S)$ satisfies the LTSS property with priority function $G(s_i) = x_i$. Moreover, $F(S)$ also satisfies the *strong LTSS* property, which guarantees that $S = \{s_{(1)} \ldots s_{(j)}\}$ is the highest-scoring subset among those subsets with cardinality $j$. The key step in this proof is to show that, if there exist two locations $s_{\text{in}} \in S$ and $s_{\text{out}} \notin S$, where $G(s_{\text{in}}) \leq G(s_{\text{out}})$, then $F(S) \leq F(S \cup \{s_{\text{out}}\} \setminus \{s_{\text{in}}\})$, i.e., the score of subset $S$ will be increased by substituting the higher-priority element $s_{\text{out}}$ for the lower-priority element $s_{\text{in}}$. This step follows from the monotonicity of function $F$. As a corollary, we can show that a large class of nonparametric scan statistics, which compare the actual and expected numbers of p-values in subset $S$ that are significant at level $\alpha$, satisfy LTSS and strong LTSS. See McFowland III et al. (2013) for details. Functions satisfying strong LTSS allow some useful optimization approaches that functions which only satisfy (weak) LTSS do not, such as efficient constrained optimization over subsets with hard constraints on region density. However, most commonly used scan statistics only satisfy the weak but not strong LTSS property. Nevertheless, the weak property is sufficient for efficient optimization over subsets of the data.
- Let $F(S)$ be an expectation-based scan statistic for any distribution in the exponential family. Then $F(S)$ satisfies LTSS with priority function $G(s_i) = q_{\max}(x_i, \mu_i)$, where $x_i$ is the observed value at location $s_i$, $\mu_i$ is the expected value at location $s_i$, and $q_{\max}(x_i, \mu_i)$ is the value $q > 1$ such that $F(S \mid q) = 0$ for $S = \{s_i\}$. See Speakman et al. (2014b) for details. As a corollary, the commonly used expectation-based Poisson and Gaussian scan statistics satisfy LTSS, as do the expectation-based exponential, binomial, and negative binomial. In earlier work, Neill (2012) proved that all expectation-based scan statistics in the *separable exponential family*, a subfamily of the exponential family which contains the Poisson, Gaussian, and exponential distributions but not the binomial and negative binomial, satisfy LTSS, with the simpler (and easier to compute) priority function $G(s_i) = \frac{x_i}{\mu_i}$. However, Speakman et al. (2014b) present a counterexample showing that the expectation-based binomial does not satisfy LTSS with this priority function.

### Incorporating Constraints into Fast Subset Scanning

Since the LTSS property only guarantees an exact solution to the unconstrained (all subsets) optimization problem, the biggest challenge within the fast subset scanning framework is to incorporate real-world constraints such as spatial proximity, graph connectivity, and temporal consistency to ensure that relevant and useful subsets are detected. A number of recent extensions use the unconstrained fast subset scan as a building block to develop powerful methods for constrained optimization. The original work on fast subset scanning (Neill 2012) demonstrated how spatial proximity constraints can be incorporated, using spatial information to constrain the search by penalizing or excluding unlikely subsets (e.g., spatially dispersed or highly irregular regions). The *fast localized scan* approach (Neill 2012) constrains the search to subsets of the local neighborhoods formed by considering each spatial location $s_i$ and its $k - 1$ nearest neighbors, for

some fixed neighborhood size $k$. Fast localized scan performs a separate, unconstrained search over subsets for each of the $N$ neighborhoods formed in this way, reducing the computational complexity of searching each neighborhood from $O(2^k)$ to $O(k)$ using the LTSS property. Thus, the overall complexity is reduced from exponential to $O(Nk + N \log N)$, where the first term describes the complexity of searching over the $N$ neighborhoods and the second term corresponds to the initial step of sorting the $N$ locations by priority. If a good choice of neighborhood size $k$ is not known, an alternative is the *fast multiscan* (Neill 2012), which compares the penalized scores $\max_S F(S \mid k) - \lambda k$ for all neighborhood sizes $k = 1 \ldots N$ and some constant $\lambda > 0$. Given labeled training data, the value of $k$ for fast localized scan or the value of $\lambda$ for fast multiscan can be chosen by cross-validation. Neill (2012) examined the detection power and spatial accuracy of the fast localized scan and fast multiscan approaches as compared to the traditional Kulldorff's spatial scan (searching over circular regions), using simulated disease outbreaks injected into real-world hospital emergency department data. The proximity-constrained subset scans substantially improved the timeliness and accuracy of detection, detecting 2 days faster with fewer than half as many missed outbreaks.

For the extension of linear-time subset scanning to graph or network data, we monitor one or more data streams at each node of the graph and wish to detect the most anomalous subset of nodes subject to the graph connectivity constraints (i.e., the given subset of nodes must form a connected subgraph of the original graph). For spatial data, the graph edges could represent spatial adjacency or travel patterns, but this framework also enables analysis of nonspatial network data. As noted above, exact optimization of the score function $F(S)$ over connected subgraphs is difficult: the FlexScan approach (Tango and Takahashi 2005) performs an exhaustive search and thus does not scale beyond 25 or 30 nodes, while other approaches (Patil and Taillie 2004; Duczmal and Assuncao 2004; Duczmal et al. 2007; Costa et al. 2012) are not guaranteed to find the highest-scoring subgraph. While the highest-scoring subset found by unconstrained LTSS may be disconnected, making it challenging to apply LTSS directly for optimization with connectivity constraints, an alternate formulation of the LTSS property can be used to speed up the search. As noted above, in the unconstrained case, we can prove that a subset $S$ is suboptimal if there exist locations $s_{\text{in}} \in S$ and $s_{\text{out}} \notin S$ where $G(s_{\text{in}}) \leq G(S_{\text{out}})$. When optimizing over connected subgraphs instead of all subsets, a variant of this property still applies, but subgraph $S$ is only provably suboptimal if the resulting subgraphs $S \cup \{s_{\text{out}}\}$ and $S \setminus \{s_{\text{in}}\}$ remain connected.

This property was recently incorporated into a depth-first search procedure, the GraphScan algorithm (Speakman et al. 2014a). By identifying and pruning paths that are provably suboptimal, GraphScan can rule out large numbers of subsets without evaluating each one individually. This approach dramatically reduces the size of the search space and the resulting computation time. Additional speed improvements are obtained by *branch and bounding*, using the unconstrained maximum score (which is efficiently computable using LTSS) as an upper bound on the maximum score of connected subgraphs and ruling out large numbers of subsets with upper bounds less than the highest-scoring subgraph found so far. See Speakman et al. (2014a) for details. The resulting GraphScan algorithm, like FlexScan (Tango and Takahashi 2005), still requires exponential computation time in the worst case, but it scales to graphs an order of magnitude larger than FlexScan, with a 450,000x speedup for graphs of size 30 (Speakman et al. 2014a). Moreover, GraphScan still identifies the highest-scoring subgraph: it is an exact, rather than approximate, algorithm. An alternative approach, Additive GraphScan (Speakman et al. 2013), can only be used for additive score functions and is not guaranteed to find the highest-scoring subgraph. However, Additive GraphScan can scale to graphs with tens of thousands of nodes and identifies near-optimal subsets with high probability in practice (Speakman et al. 2013).

Another recent extension of the fast subset scanning framework, the Dynamic Subset Scan (Speakman et al. 2013), focuses on

detecting *dynamic* patterns, where the affected subset of locations can grow, shrink, or move over time. Typical space-time scan approaches (Kulldorff 2001; Neill et al. 2005) search over space-time cylinders, assuming that the affected subset of locations remains constant for the duration of the event. However, this over-constrained approach leads to reduced detection power for dynamically evolving events, as well as failing to accurately capture the event dynamics. An alternative, under-constrained approach of performing independent spatial scans for each time step results in identified patterns that display unrealistic temporal trends (e.g., affecting the east side of the city on day 1, the west side on day 2, and back to the east side on day 3). Thus, the Dynamic Subset Scan optimizes the score function $F(S)$ over subsets of locations at each time step while enforcing *temporal consistency constraints*, considering the patterns detected at adjacent time steps, and rewarding patterns that are not dramatically different between time steps $t$ and $t + 1$. This approach allows the spatial extent of an event to evolve smoothly over time while penalizing unrealistic event dynamics.

Unlike the fast subset scanning approaches described above, which enforce *hard constraints* and thus rule out some subsets from consideration, Dynamic Subset Scan enforces *soft constraints*, which can be interpreted as applying bonuses or penalties to the score function for including or excluding certain locations. This is a specific case of the more general, *penalized fast subset scanning* (PFSS) framework described by Speakman et al. (2014b). Incorporating penalties is difficult because a penalized version of the score function may not satisfy the LTSS property. However, Speakman et al. (2014b) show that any expectation-based scan statistic in the exponential family can be written as an additive function conditional on the relative risk parameter $q$. Only a linear number of ranges for $q$ must be considered, and optimization over subsets for each $q$ range is very efficient. Moreover, this formulation allows bonuses or penalties for each location to be incorporated into the score function while maintaining efficient optimization. In the Dynamic Subset Scan, PFSS is used to optimize over all subsets of locations for a given time step, with location-specific bonuses or penalties based on the detected subsets for the previous and next time steps, and incorporating a flexible, generative model of event propagation. This efficient conditional optimization step is iterated until convergence, thus propagating information both backward and forward in time. Connectivity constraints can also be incorporated into the Dynamic Subset Scan framework, requiring the use of Additive GraphScan (rather than simply including all locations that make a positive contribution to the score) for each step. Speakman et al. (2013) applied the Dynamic Subset Scan (with connectivity and temporal consistency constraints) to detection, tracking, and source tracing of spreading contaminants in a water distribution network. Dynamic Subset Scan demonstrated earlier detection of contamination events and more accurate identification of the affected subset of nodes through time.

### Multivariate Fast Subset Scanning

While the fast subset scan approaches described above focus on the univariate case, monitoring a single spatiotemporal data stream, these approaches can also be extended to the multivariate case. The multivariate fast subset scan (Neill et al. 2013) can be used to monitor multiple streams of space-time data, identifying subsets of streams where the recently observed counts are significantly higher than expected. Similarly, the Fast Generalized Subset Scan (FGSS) can be used to discover patterns in general multivariate datasets, identifying subsets of similar data records with anomalous values for some subset of attributes (McFowland III et al. 2013). The key idea for both approaches is similar: linear-time subset scanning can be used for efficient optimization over subsets of locations (or records) for a given subset of streams (or attributes) but can also be used for efficient optimization over subsets of streams (or attributes) for a given subset of locations (or records). Thus, we can iterate between these two efficient conditional optimization steps until a local maximum of the score function is reached and perform multiple restarts in order to approach the global maximum.

The multivariate fast subset scan builds on the univariate fast subset scanning approach, jointly optimizing a parametric log-likelihood ratio statistic (such as the expectation-based Poisson statistic described above) over proximity-constrained subsets of locations and over all subsets of data streams. The most natural formulation of the multivariate scan statistic in this setting, *subset aggregation*, assumes a constant multiplicative increase across all affected streams and thus adds counts and baselines across the monitored subset of streams. An alternative formulation by Kulldorff et al. (2007) proposes adding log-likelihood ratios across streams (assuming that the data streams are conditionally independent). Neill et al. (2013) demonstrate that the Kulldorff's multivariate scan can also be made efficient using the LTSS property, by iterating between two steps: optimizing over subsets of records (for given values of the multiplicative effect of the event on each data stream) and recalculating the maximum likelihood values of the event's effects for the given subset of records. Regardless of which formulation is used, the multivariate fast subset scan is computationally efficient and scales to large numbers of locations and streams. Moreover, significant gains in detection power and spatial accuracy were observed when searching over subsets of data streams and when detecting proximity-constrained subsets of locations rather than searching over circular regions (Neill et al. 2013).

The FGSS approach (McFowland III et al. 2013) does not assume space-time data but instead considers an arbitrary set of attributes measured for each of a large set of data records. Nonparametric scan statistics are used to convert the disparate attributes to the same scale (empirical p-values between 0 and 1) and to integrate these values in a principled statistical framework. FGSS consists of four steps: (1) efficiently learning a Bayesian network model that represents the assumed null distribution of the data; (2) computing the conditional probability of each attribute value in the dataset given the Bayes Net, conditioned on the other attribute values for that record; (3) computing an empirical p-value range corresponding to each attribute value by ranking the conditional probabilities, where under the null hypothesis we expect empirical p-values to be uniformly distributed on [0,1]; and (4) using a nonparametric scan statistic to detect subsets of records and attributes with an unexpectedly large number of low (significant) empirical p-values. The final step is computationally expensive (exponential in the numbers of records and attributes for a naive search), but LTSS can be used to speed up this search, converging to a local maximum of the score function and ensuring that each iteration step is linear (not exponential) in the number of records or attributes. FGSS was shown to consistently outperform previously proposed methods in terms of detection power and characterization accuracy across multiple application domains, and scales to much larger datasets, thus enabling accurate and efficient pattern detection in massive, high-dimensional data.

## Key Applications

One important real-world application of subset scanning is in the area of disease surveillance, where we attempt to detect emerging outbreaks of disease in their very early stages by identifying anomalous clusters of disease cases. In the multivariate disease surveillance setting, we monitor a set of data streams $D_m$ ($m = 1 \ldots M$) on a regular (e.g., daily or hourly) basis at a set of spatial locations (e.g., zip codes) $s_i$ ($i = 1 \ldots N$). For each combination of location and data stream, we have a time series of observed counts $c^t_{i,m}$, where each count $c^t_{i,m}$ could represent the number of observed cases of a given type (e.g., emergency department visits with respiratory complaints) in a given zip code on a given day. A typical goal in this setting is to identify spatial regions (subsets of locations) where some subset of the monitored data streams have recent counts that are higher than expected. Here, the expected counts $b^t_{i,m}$ are obtained through time series analysis of historical data and can account for trends such as the day of week, seasonality, holidays, and known events. Multiple variants of subset

scanning have been evaluated on the disease surveillance task, typically through semisynthetic testing (in which simulated outbreaks are injected into real-world background data). Searching over proximity-constrained subsets of locations to identify irregularly shaped spatial clusters (Neill 2012) enables earlier and more accurate outbreak detection, as measured by the average number of days to detect and proportion of outbreaks detected for a given false-positive rate, as well as the spatial overlap between true and identified outbreak regions. Further improvements in detection power can be obtained by integrating information from multiple health data streams (Neill et al. 2013), searching over subsets of streams as well as proximity-constrained subsets of locations. This approach also helps to characterize outbreaks by identifying the affected subset of streams, and a further extension, the "multidimensional subset scan" (Neill and Kumar 2013) can also identify differentially affected subpopulations (e.g., by gender, age, socioeconomic status, or behavioral risk factors).

Incorporating other constraints into the subset scan, such as graph connectivity and temporal consistency (Speakman et al. 2013, 2014a) or similarity between records in general datasets (McFowland III et al. 2013), enables a wide variety of other applications to be addressed using this framework. For example, Speakman et al. (2013) demonstrate improved performance for detecting, tracking, and source tracing contamination events spreading through a water distribution system. McFowland III et al. (2013) evaluate their approach on outbreak detection, customs monitoring of container shipments, and computer network intrusion detection, demonstrating improvements over the current state of the art in all three application domains. Finally, recent extensions of subset scanning have been applied to detect patterns in massive, complex real-world datasets such as images (Somanchi and Neill 2013), text (Nobles et al. 2014), and online social networks such as Twitter (Chen and Neill 2014). Somanchi and Neill (2013) demonstrate that their approach can be used to accurately detect prostate cancer and identify other regions of potential interest in digital pathology slides. Nobles et al. (2014) apply the subset scan approach to identifying emerging "novel" disease outbreaks with previously unseen or anomalous patterns of symptoms, using free-text emergency department chief complaint data. Chen and Neill (2014) developed a new approach to event detection in heterogeneous social media graphs and applied this approach to advance prediction of civil unrest events (strikes, protests, and riots) and early warning for rare disease outbreaks (hantavirus), using Twitter data from Latin America. In both application domains, their approach outperformed five competing, state-of-the-art methods for both event detection and forecasting, increasing detection power, forecasting accuracy, and forecasting lead time while reducing time to detection (Chen and Neill 2014).

## Future Directions

While subset scanning is a rapidly emerging and highly promising field, a number of challenging open problems remain to be addressed. One avenue for future research is continuing to extend the range of score functions for which the linear-time subset scanning property can be proven to hold, as well as the range of constraints that can be incorporated into the fast subset scan framework while still allowing computationally efficient and scalable solutions. These extensions have the potential to expand the use of subset scanning for a variety of real-world applications requiring analysis of massive, complex datasets. A second important direction is gaining a better understanding of the statistical properties of subset scanning, for example, identifying necessary and sufficient conditions for which the highest-scoring subset converges to the true affected subset or a provably good approximation or quantifying the detection power of constrained subset scans as a function of how well the chosen constraints correspond to the true pattern of interest. Finally, in many cases, event detection can be thought of as integrating information from

many noisy sensors. This *sensor fusion* problem, assuming a given set of noisy sensors, complements the *sensor placement* problem, in which sensors are often assumed to be perfect and the focus is on optimally placing sensors in space or on a network. Another useful property, submodularity, can be used to efficiently find near-optimal solutions to a variety of sensor placement problems (Leskovec et al. 2007), and it is an open problem whether the submodularity and linear-time subset scanning properties can be effectively combined to solve problems requiring both placement of, and integration of data from, noisy sensors. One example application where this might be useful is in the crowdsourced collection of environmental and ecological data (e.g., observations of plant and animal species or measurement of air, water, and soil quality) by "citizen scientists." In this case, data quality varies considerably based on individuals' expertise, and the accuracy of the data might be substantially improved by asking individuals with relevant expertise if they are willing to perform analyses of specific types or in specific locations.

## Cross -References

- Hotspot Detection, Prioritization, and Security
- Irregular Shaped Spatial Clusters: Detection and Inference
- Linear Anomalous Window
- Movement Patterns in Spatio-Temporal Data
- Public Health and Spatial Modeling

## References


Chen F, Neill DB (2014) Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In: Proceedings of the 20th ACM SIGKDD conference on knowledge discovery and data mining, New York, pp 1166–1175

Costa MA, Assuncao RM, Kulldorff M (2012) Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Comput Stat Data Anal 56(6):1771–1783

Das K, Schneider J, Neill DB (2008) Anomaly pattern detection in categorical datasets. In: Proceedings of the 14th ACM SIGKDD conference on knowledge discovery and data mining, Las Vegas, pp 169–176

Duczmal L, Assuncao R (2004) A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. Comput Stat Data Anal 45:269–286

Duczmal L, Cancado A, Takahashi R, Bessegato L (2007) A genetic algorithmic for irregularly shaped scan statistics. Comput Stat Data Anal 52(1):43–52

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26(6):1481–1496

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

Kulldorff M, Huang L, Pickle L, Ducmzal L (2006) An elliptic spatial scan statistic. Stat Med 25:3929–3943

Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R (2007) Multivariate scan statistics for disease surveillance. Stat Med 26:1824–1833

Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD conference on knowledge discovery and data mining, San Jose, pp 420–429

McFowland III E, Speakman S, Neill DB (2013) Fast generalized subset scan for anomalous pattern detection. J Mach Learn Res 14:1533–1561

Naus JI (1965) The distribution of the size of the maximum cluster of points on the line. J Am Stat Assoc 60:532–538

Neill DB (2009a) An empirical comparison of spatial scan statistics for outbreak detection. Int J Health Geogr 8:20

Neill DB (2009b) Expectation-based scan statistics for monitoring spatial time series data. Int J Forecast 25:498–517

Neill DB (2012) Fast subset scan for spatial pattern detection. J R Stat Soc Ser B Stat Methodol 74(2):337–360

Neill DB, Kumar T (2013) Fast multidimensional subset scan for outbreak detection and characterization. Online J Publ Health Inf 5(1):156

Neill DB, Moore AW (2004) Rapid detection of significant spatial clusters. In: Proceedings of the 10th ACM SIGKDD conference on knowledge discovery and data mining, Seattle, pp 256–265

Neill DB, Moore AW, Sabhnani M, Daniel K (2005) Detection of emerging space-time clusters. In: Proceedings of the 11th ACM SIGKDD conference on knowledge discovery and data mining, Chicago, pp 218–227

Neill DB, McFowland III E, Zheng H (2013) Fast subset scan for multivariate event detection. Stat Med 32:2185–2208

Nobles M, Deyneka L, Ising A, Neill DB (2015) Identifying emerging novel outbreaks in textual emergency department data. Online J Publ Health Inf 7(1): e45

Patil GP, Taillie C (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ Ecol Stat 11:183–197

Somanchi S, Neill DB (2013) Discovering anomalous patterns in large digital pathology images. In: Proceedings of the 8th INFORMS workshop on data mining and health informatics, Minneapolis

Speakman S, Zhang Y, Neill DB (2013) Dynamic pattern detection with temporal consistency and connectivity constraints. In: Proceedings of the 13th IEEE international conference on data mining, Dallas, pp 697–706
Speakman S, McFowland III E, Neill DB (2015) Scalable detection of anomalous patterns with connectivity constraints. J Comput Graph Stat 24(4):1014–1033
Speakman S, Somanchi S, McFowland III E, Neill DB (2016, in press) Penalized fast subset scanning. J Comput Graph Stat
Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4:11

## Summary Information

▸ Metadata and Interoperability, Geospatial

## Supplementary Material

▸ Metadata and Interoperability, Geospatial

## Surface Modeling

▸ Aggregate Data: Geostatistical Solutions for Reconstructing Attribute Surfaces

## Surveillance

▸ Data Collection, Reliable Real-Time
▸ Evolution of Earth Observation

## Survey Knowledge

▸ Wayfinding, Landmarks

## Susceptibility Analysis

▸ Sensitivity Analysis

## Sustainability Risk

▸ Climate Risk Analysis for Financial Institutions

## Sustainable Development

▸ Climate Change and Developmental Economies

## SVG

▸ Scalable Vector Graphics (SVG)
▸ Web Mapping and Web Cartography

## Sweep Line Algorithm

▸ Plane Sweep Algorithm

## Synchronization of Spatial Data

▸ Positional Accuracy Improvement (PAI)

## Synonymy

▸ Retrieval Algorithms, Spatial