Fast and Flexible Outbreak Detection by Linear-Time Subset Scanning

Daniel B. Neill, Ph.D.

Heinz School of Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

We present a new method of “linear-time subset scanning” and apply this technique to various spatial outbreak detection scenarios, making it computationally feasible (and very fast) to perform spatial scans over huge numbers of search regions.

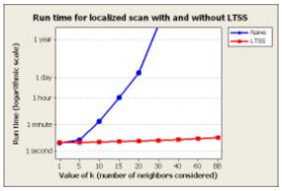
BACKGROUND

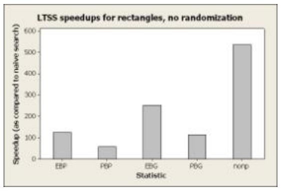
The spatial scan statistic [1] detects significant spatial clusters of disease by maximizing a likelihood ratio statistic over a large set of spatial regions. Typical spatial scan approaches either constrain the search regions to a given shape, reducing power to detect patterns that do not correspond to this shape, or perform a heuristic search over a larger set of irregular regions, in which case they may not find the most relevant clusters. In either case, computation time is a serious issue when searching over complex region shapes or when analyzing a large amount of data. An alternative approach might be to search over all possible subsets of the data to find the most relevant patterns, but since there are exponentially many subsets, an exhaustive search is computationally infeasible.

METHODS

We prove that many commonly used scan statistics (including the Poisson [1-2], Gaussian [2], and nonparametric [3] statistics) have the property of “linear-time subset scanning” (LTSS): the subset of locations which maximizes the likelihood ratio can be found by ordering the locations according to some relevance criterion and searching over groups consisting of the top-k most relevant records, requiring linear rather than exponential time. For parametric statistics, we sort by the ratio of observed to expected count. If all subsets of the data are equally likely to be affected (e.g. outbreaks that do not cluster spatially), we apply LTSS directly to find the most significant subset of N locations, searching only N regions instead of 2^N.

However, we often want to use spatial information to constrain our search by penalizing or excluding unlikely subsets (e.g. disconnected or highly irregular regions). In such cases, LTSS cannot be used directly to find the optimal subset subject to these constraints, but it can speed up our search in several ways. First, for some sets of search regions, we can find the optimal region using multiple LTSS searches. For example, we can search over regions consisting of a center location and some (not necessarily connected) subset of its k-nearest neighbors. LTSS reduces the complexity of this “localized scan” (similar to FlexScan [4]) from exponential to linear in k. Second, we can use the unconstrained maximum score as an upper bound on the constrained maximum score, performing a branch-and-bound search. For example, we can search all distinct rectangular regions using LTSS in a new variant of the “fast spatial scan” [5].





RESULTS

We compared the computation time needed to perform spatial scans, with and without LTSS, on 281 days of ED visit data from 88 Allegheny County zip codes. Various scan statistics and sets of search regions were considered. Our main results include:

1) LTSS scans over all distinct subsets in 3.65 seconds. Naïve search would require ~10^{25} years.

2) LTSS performs localized scans in 4-8 seconds for all k. Naïve search is infeasible for k > 20 (Figure 1).

3) LTSS scans over all distinct rectangles 57-534x faster than naïve search, requiring between 16 seconds and 2 minutes vs. over 2 hours (Figure 2).

CONCLUSIONS

LTSS is a powerful and useful tool that enables us to speed up a wide variety of spatial outbreak detection methods. We are currently extending this method to the Bayesian, space-time, and multivariate scan statistics, and these extensions will make an even wider range of searches computationally feasible and fast.

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8-R01-HK000020-02.

REFERENCES

[1] Kulldorff M, A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997, 26(6): 1481-1496.

[2] Neill DB, Detection of Spatial and Spatio-Temporal Clusters. Ph.D. thesis, Carnegie Mellon University, 2006.

[3] Neill DB, Lingwall J, A nonparametric scan statistic for multivariate disease surveillance. Adv Disease Surv, 2007, 4: 106.

[4] Tango T, Takahashi K, A flexibly shaped spatial scan statistic for detecting clusters. Intl J Health Geographics, 2005, 4: 11.

[5] Neill DB, Moore AW, Rapid detection of significant spatial clusters. Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2004, 256-265.

Further Information:
Daniel B. Neill, neill@cs.cmu.edu
http://www.cs.cmu.edu/~neill