An Expectation-Based Scan Statistic for Detection of Space-Time Clusters

Daniel B. Neill, Andrew W. Moore, Maheshkumar R. Sabhnani, Kenny Daniel
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

This paper describes a new class of space-time scan statistics designed for rapid detection of emerging disease clusters. We evaluate these methods on the task of prospective disease surveillance, and show that our methods consistently outperform the standard space-time scan statistic approach.

BACKGROUND

The space-time scan statistic [1] is a powerful statistical tool for prospective disease surveillance. It searches over a set of spatio-temporal regions (each representing some spatial area S for the last k days), finding the most significant regions (S, k) by maximizing a likelihood ratio statistic, and computing p-values of these potential clusters by randomization.

The standard, "population-based" method assumes that, for each spatial location s_i on each day t, we have a *population* p^t_i and a *count* (observed number of cases) c^t_i. Then, under the null hypothesis of no clusters, we expect each count c^t_i to be proportional to its population p^t_i. We then search for regions (S, k) with disease rate (cases per unit population) significantly higher inside the region than outside. In the original space-time scan statistic [1], the populations are assumed to be given, and in [2], populations are estimated assuming independence of space and time.

Here we propose an alternative, "expectation-based" method, in which we infer the *expected number of cases* b^t_i in each spatial location, based on the time series of previous counts. In this case, under the null hypothesis of no clusters, we expect each count c^t_i to be equal to b^t_i, rather than proportional to population. We then search for regions (S, k) with counts that are significantly higher than expected.

METHODS

Our expectation-based method consists of two parts: first, we infer the expected counts for recent days from the time series of past counts in each location, accounting for day of week and seasonal trends, using one of the time series analysis methods given in [3]. Second, we use a new "emerging cluster" scan statistic to detect regions with recent counts significantly higher than expected, assuming that the relative risk in a disease cluster increases monotonically over time. This is different than the standard, "persistent cluster" approach, which assumes constant relative risk over the outbreak's duration.

To test detection power, we compared the various space-time scan statistics (population-based vs. expectation-based, emerging vs. persistent) on eight types of simulated respiratory outbreaks, injected into real (anonymized) Emergency Department records and over-the-counter sales data for Allegheny County, PA. These included simulated anthrax outbreaks generated by the BARD simulator [4], as well as simpler outbreaks with linear onsets, each with various parameter settings. We also examined the effects of temporal window size, level of aggregation, and time series analysis method on detection power.

RESULTS

On the outbreaks tested, the expectation-based statistic consistently outperformed the population-based statistic, detecting a higher proportion of outbreaks and requiring fewer days to detect, at a false positive rate of 1/month. Similarly, the emerging cluster statistic consistently outperformed the persistent cluster statistic. Simple time series analysis methods such as "mean of last 28 days" performed well on ED data, while more complicated time series methods such as exponentially weighted linear regression (EWLR) and a method based on [2] performed better for the OTC data. Optimal temporal window size varied from 1-7 days, with longer windows preferable for more slowly growing outbreaks.

CONCLUSIONS

We have presented a new class of space-time scan statistics for detection of emerging clusters, and demonstrated that these methods are highly successful on the task of rapidly and accurately detecting emerging disease outbreaks. More details of our methods and results are given in the full paper [3].

REFERENCES

[1] Kulldorff M, Prospective time-periodic geographical surveillance using a scan statistic. Journal of the Royal Statistical Society A, 2001, 164: 61-72.

[2] Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F, A space-time permutation statistic for the early detection of disease outbreaks. PLoS Medicine, 2005, 2(3):e59.

[3] Neill DB, Moore AW, Sabhnani MR, Daniel K, Detection of emerging space-time clusters. Proc. 11th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2005, 218-227.

[4] Hogan W, Cooper G, Wagner M, Wallstrom G, A Bayesian anthrax aerosol release detector. Technical Report, RODS Laboratory, University of Pittsburgh, 2004.

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