A Nonparametric Scan Statistic for Multivariate Disease Surveillance

Daniel B. Neill, Ph.D., Jeff Lingwall, M.S.

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

OBJECTIVE

We present a new method for multivariate outbreak detection, the "nonparametric scan statistic" (NPSS). NPSS enables fast and accurate detection of emerging space-time clusters using multiple disparate data streams, including nontraditional data sources where standard parametric model assumptions are incorrect.

BACKGROUND

Expectation-based scan statistics [1] extend the traditional spatial and space-time scan statistics [2-3] by inferring expected counts for each location from past data and detecting regions where recent counts are higher than expected. While these methods have been shown to achieve high detection power across a variety of datasets and outbreak types [4], they make strong parametric model assumptions (e.g. Poisson or Gaussian counts), and performance degrades when these models are incorrect. Our solution, NPSS, is a new scan statistic approach which does not assume a parametric model, but instead combines empirical p-values across multiple locations, days, and data streams to discover significant disease clusters.

METHODS

Given a time series of observed counts $c_{i,m}^t$ for each data stream D_m at each location s_i, we wish to detect spatial regions where the recent counts for some subset of streams are higher than expected. To do so, we first compute empirical p-values $P_{i,m}^t$ for each stream and location for each recent day: these are defined as $(T_{beat}+ 1) / (T+1)$, where T_{beat} is the number of past days (for that data stream and location) with higher residuals and T is the total number of past days. We then scan over the space-time regions (D, S, W), each consisting of some subset of streams D for some spatial area S for the last W days. We search for regions where the $P_{i,m}^t$ are significantly lower than expected, corresponding to higher than expected counts.

Under assumptions of independence and stationarity, we expect each empirical p-value to be asymptotically uniformly distributed on [0, 1] under the null hypothesis of no outbreaks. Thus for a given region (D, S, W) with N empirical p-values (N is the product of the numbers of locations, streams, and days), we expect the number of locations significant at level α to be binomially distributed with parameters N and α. Following [5], we define a region's score F(D, S, W) to be $\max_\alpha (N_\alpha - N\alpha) / (N\alpha(1-\alpha))^{1/2}$, where N_α is the number of empirical p-values less than α. We can find the most significant region by maximizing F(D, S, W) over all D, S, and W, and calculate the statistical significance of this region by randomization testing, generating empirical p-values uniformly on [0,1].

RESULTS

We first compared the univariate NPSS method to the expectation-based Poisson (EBP) scan statistic on simulated outbreaks injected into five Allegheny County datasets: respiratory ED visits, three streams of OTC data (cough/cold, antifever, thermometers), and simulated binary biological sensor data. At a fixed false positive rate of 1/month, NPSS performed comparably to EBP for the four traditional data sources (detecting 0.5 days faster for ED and 0.3 days slower for OTC data), but also detected 0.7 days faster for the sensor data. The multivariate NPSS method showed further gains in detection power for outbreaks that simultaneously affected the multiple streams of OTC data, detecting in 3.71 days at 1 fp/month, as compared to 4.12, 4.81, and 5.34 days respectively for the three univariate NPSS detectors.

CONCLUSIONS

NPSS increases detection power by combining evidence from multiple disparate data streams, and outperforms other scan statistic approaches on datasets (such as binary sensor data) where typical parametric models are incorrect. Unlike our recently proposed multivariate Bayesian scan statistic [6], NPSS does not model and differentiate between different outbreak types. Instead, it is a general multivariate anomaly detector that can detect emerging outbreaks and identify the affected locations and data streams.

This work was partially supported by NSF grant IIS-0325581 and CDC grant 1 R01 PH000028-01.

REFERENCES

[1] Neill DB, Moore AW, Methods for detecting spatial and spatio-temporal clusters. Handbook of Biosurveillance, 2006, 243-254.

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

[3] Kulldorff M, Prospective time-periodic geographical surveillance using a scan statistic. J Royal Stat Soc A, 2001, 164: 61-72.

[4] Neill DB, An empirical comparison of spatial scan statistics for outbreak detection. Submitted for publication.

[5] Donoho D, Jin J, Higher criticism for detecting sparse heterogeneous mixtures. Annals of Statistics 2004, 32(3): 962-994.

[6] Neill DB, Moore AW, Cooper GF, A multivariate Bayesian scan statistic. Advances in Disease Surveillance 2007, 2: 60.

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