Ellipse-based Clustering Analysis Using a Time Series Algorithm

Fu-Chiang Tsui, Ph.D., Jialan Que, M.S., Michael M. Wagner, M.D., Ph.D.

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh

OBJECTIVE

This paper describes a new clustering algorithm Ellipse-based Clustering Analysis (ECA), which uses a time series algorithm to identify zip codes with abnormal counts, and uses a pattern recognition method to identify spatial clusters in ellipse shapes. Using ellipses could help detect elongated clusters resulting from wind dispersion of bio-agents. We applied the ECA to over-the-counter (OTC) medicine sales. The pilot study demonstrated the potential use of the algorithm in detection of clustered outbreak regions that could be associated with aerosol release of bio-agents.

BACKGROUND

Many cities in the US and the Center for Disease Control and Prevention (CDC) have deployed biosurveillance systems to monitor regional health status. Biosurveillance systems rely on algorithms[1] that analyze data in temporal domain (*e.g.,* CuSUM) and/or spatial domain (*e.g.,* SaTScan). Spatial domain-based algorithms often require population information to normalize the counts (*e.g.,* emergency department visits) within a geographic region. This paper presents a new algorithm ECA that analyzes data in both temporal and spatial domains--using time series analysis for each of zip codes with abnormal counts and using pattern recognition methods for spatial clusters.

METHODS

The National Retail Data Monitor (NRDM) collects daily OTC sales from 20,500+ retail stores.[2] NRDM employs advanced schemes for fast multi-year data retrieval.[2] ECA first ran a time series algorithm—Wavelet anomaly detector (WAV)[3]—that retrieved one year worth of historical data for computing each day's expected counts between May 8 to May 23, 2006 (the study period) for each zip code in PA and identified a subset of "hot" zip codes with increased sales compared with the expected counts. The ECA then retrieved the hot zip codes and identified ellipse clusters using modified K-means method and cluster validity indexes. Within each initially identified cluster, the ECA searches sub-clusters based on the criteria of connected hot zip codes and then determines top 5 clusters using maximum log likelihood ratio (MLLR). We also define *computational time* as the time of processing one-day OTC data starting from one-year baseline query to the search of top 5 clusters.

RESULTS

WAV analyzed total 494 zip codes with sales data in PA for OTC product category *anti-fever adult* medicine each day during the study period. Figure 1 shows the top 5 clusters (enclosed by ellipses in light yellow) for May 8, 2006 data and the small irregular shapes in dark yellow represents hot zip codes that constitute a small fraction of total 494 zip codes. The average computational time for one-day OTC data during the study period was about 2.5 minutes.

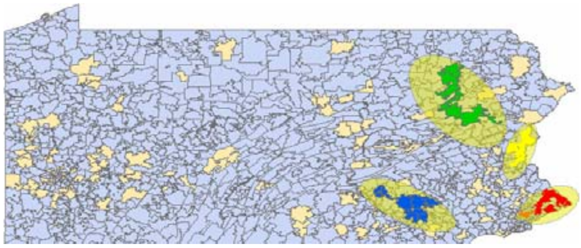

Figure 1 – Map of Pennsylvania with 5 detected elliptic clusters of *anti-fever-adult* over-the-counter medicine sales, identified by ECA algorithm, for May 8, 2006. Each irregular shape in dark yellow represents a "hot" zip code with elevated sales. The 5 ellipses contain hot zip codes in red, orange, yellow, blue, and green. The ellipse close to Philadelphia with red zip copes has MLLR.

DISCUSSION

The limit of the study is the lack of randomization test and rigorous evaluation. We are in process of comparing ECA with SaTScan that performs exhaustive search of elliptic clusters with MLLR. We also plan to use semi-synthetic data generated by the Bayesian Aerosol Release Detection algorithm that injects cases into an area based on wind dispersion model. Nevertheless, the contribution of this paper is to introduce a new algorithm and demonstrated its pilot results. ECA has the following advantages: 1) use of ellipses that have the advantage of detecting elongated clusters resulting from wind dispersion of bio-agents, 2) use of time series analysis for extraction of hot zip codes without population information, and 3) use of a small fraction of zip codes for fast computational time.

ACKNOWLEDGEMENT

This material is based upon work supported by grants NSF-0325581, CDC-5R01PH00026-02, NLM-5R21LM008278-03, PA-DOH-ME-01737, and AFRL-F30602-01-2-0550.

REFERENCES

1. Wagner M, Moore A, Aryel R. Handbook of Biosurveillance: Academic Press; 2006.
2. Tsui F-C, Espino JU, et. al. Key design elements of a data utility for national biosurveillance: event-driven architecture, caching, and web service model. Proc AMIA Symp 2005.
3. Zhang J, Tsui FC, Wagner MM, Hogan WR. Detection of outbreaks from time series data using wavelet transform. Proc AMIA Symp 2003:748-52.

Further Information: Rich Tsui, tsui@cbmi.pitt.edu.