Application of Nonlinear Data Analysis Methods to Locating Disease Clusters

Linda Moniz, PhD, William Peter, PhD
Johns Hopkins University Applied Physics Laboratory

Objective
Cluster detection with a mechanism for reducing false alarms and increasing sensitivity.

Background
Finding clusters or hot spots of disease over a wide spatial area is a priority of many epidemiologists and health departments. However, nearly all clustering algorithms currently used for this application are variants of only one algorithm - the Kulldorff spatial scan statistic [1]. Many of these variations are prone to false alarms when thresholds are set sensitively enough to find detections in the first few days of an outbreak, and the algorithms often have difficulty locating irregularly shaped clusters.

Methods
Our research centers on using techniques from nonlinear data analysis (Lorenz's Method of Analogues [2,3]) and information theory (Schreiber's Transfer Entropy [4]) together with the ACE clustering algorithm [5] to find these clusters of disease. The ACE algorithm is a particle-mesh heuristic method that has been shown to be both faster and more accurate than other clustering algorithms. We test our algorithm with geographically distributed data from both sparsely and densely populated areas and also compare its performance to a state-of-the-art scan statistic (An implementation using circular clusters, stratification by day of week, and Gumbel p-values to reduce the number of Monte Carlo runs) using specifically targeted artificially injected disease clusters on top of the real data.

Results
Although detection results with the Method of Analogues were more accurate with a longer baseline, with a minimum 60-day baseline we were able to detect anomalies that the scan statistic was not able to detect. The injected cluster was not detectable in 4/5 instances by scan statistics, but was both detected and verified to be a cluster (detection plus information transfer between spatial locations) using the nonlinear methods. Moreover, the nonlinear methods were able to verify that some clusters reported by the scan statistics were false alarms; there was no transfer of information between the locations in that cluster.

Conclusions
The nonlinear data analysis approach shows promise as an adjunct algorithm to scan statistics. This approach can computationally verify cluster detections and reduce the number of false alerts as well as detect clusters that are not visible to scan statistics.

References
[1] M. Kulldorff, A spatial scan statistic . Communications in Statistics: Theory and Methods, 26(6): 1481-1496, 1997.
[2]Lorenz, EN Atmospheric predictability as revealed by naturally occurring analogues. J. Atmosphere Sci 26 (1969).
[3]Viboud et al Prediction of the spread of influenza epidemics by the method of analogues. American Journal of Epidemiology 158 (10) (2003).
[4]Schreiber, T. Measuring information transfer. Physical Review Letters 85(2) (2000).
[5] William Peter, John Chiochetti, Clare Giardina, New unsupervised clustering algorithm for large datasets. SIGKDD '03, August 24-27 2003, Washington, DC. ACM 1-58113-737-0/03/0008.