Mixture Likelihood Ratio Scan Statistic for Disease Outbreak Detection

Jarad B. Niemi[1], Michael D. Porter[2], and Brian J. Reich[3]

[1] *Department of Statistical Science, Duke University, Durham, NC, USA.*

[2] *SPADAC Inc., McLean, VA, USA.*

[3] *Department of Statistics, North Carolina State University, Raleigh, NC, USA.*

OBJECTIVE

This article describes the methodology and results of Team #134's submission to the 2007 ISDS Technical Contest.

BACKGROUND

The prospective disease surveillance contest consisted of three synthetic outbreaks (*E. Coli*, Cryptosporidium, and Influenza) injected into three data sources (emergency department visits (ED), over-the-counter anti-diarrheal and anti-nauseant sales (OTC), and nurse hotline calls (TH)). The training data included 30 outbreak signatures for each outbreak type.

METHODS

Our method assumes a Poisson distribution for daily counts from each of the different data sources. We modeled the daily mean as a baseline plus an outbreak component. The baseline included day of week and seasonal effects. The main contribution of our approach is to explicitly model the outbreak component. Guided by the distinct outbreak signatures in the training data, we assumed parametric forms for the outbreak profiles (see Table 1) where t indicates the number of days into the outbreak. While we do not know the exact parameter values for a given outbreak, we can estimate them. This allows the development of a set of mixture likelihood ratio statistics [1,2], one for each possible outbreak starting date. The Mixture Likelihood Ratio Scan Statistic (MLRSS) is the result of scanning over the possible starting dates to find the most likely one.

Source	$\delta_t(1,\theta)$
ED	$c\exp(-[\log(t)-\mu]^2/\sigma)$
OTC	$c\exp(-[t-\mu]^2/\sigma)$
TH	$c\exp(-[(t-\mu_1)^2+(t-\mu_2)^2]/\sigma)$

Table 1: Mathematical form of the outbreak profiles.

RESULTS

The parameters for each outbreak were estimated by maximum likelihood. Figure 1 shows the result of the parametric fit for the first training outbreak.

Our contest score was 5.58 for ED, 24.00 for OTC, and 24.00 for TH. This resulted in a second place finish.

CONCLUSION

This approach, based on standard sequential change point methodologies, enjoys several properties. By taking into account the outbreak profiles, we are leveraging more information about the total process (both before and after an outbreak) than if we only considered nonspecific deviations from the "in-control" process. Furthermore, although not an objective of the contest, our method facilitates the estimation and prediction of the outbreak start time, severity, and length. This could be useful in planning and evaluating mitigation strategies once an outbreak is detected. Our method can be extended to multiple data sources, space-time surveillance, or multiple syndromes. Finally, the computation is quick enough to allow this method to be used when the data arrives much more frequently than once per day.

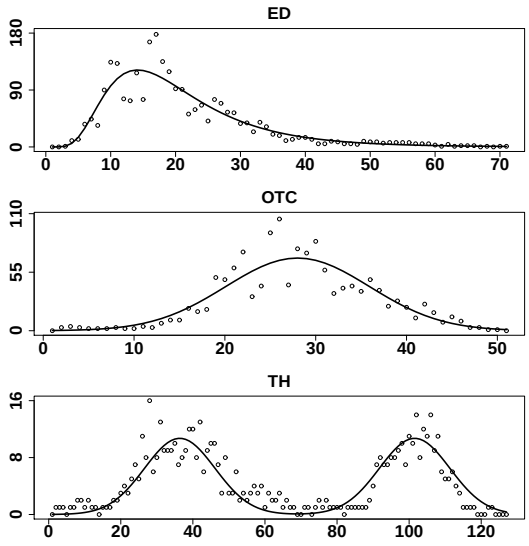

Figure 1: Outbreak profiles (*lines*) estimated for the first training outbreak (*points*) of each data source. The x-axis is days into outbreak and the y-axis is the excess counts.

REFERENCES

[1] Pollak, M. (1987) Average Run Lengths of an Optimal Method of Detecting a Change in Distribution. *Annals of Statistics 15,* 749–779.

[2] Pollak, M. and Siegmund, D. (1975) Approximations to the Expected Sample Size of Certain Sequential Tests. *Annals of Statistics 3*, 1267–1282.