Modifications to Spatial Scan Statistics for Estimated Probabilities at Fine-Resolution in Highly Skewed Spatial Distributions

James Edgerton, PhD[1], Howard Burkom, PhD[1], Michael Leuze, PhD[2]

[1]The Johns Hopkins University Applied Physics Laboratory

[2] Science Applications International Corporation

OBJECTIVE

Modifications to spatial scan statistics are investigated for prospective cluster detection at fine-resolution with highly skewed spatial distributions having many spatial zones with very few cases. Several alternative methods for the estimation of spatial probabilities and expected counts from counts in a baseline data window are evaluated with the Poisson spatial scan statistic [1] and the space-time permutation scan statistic [2] using goodness-of-fit statistics and cluster rates to compare performance.

BACKGROUND

Estimation of representative spatial probabilities and expected counts from baseline data can cause problems in applying spatial scan statistics when observed events are sparse in a large percentage of the spatial zones (e.g., zip codes or census tracts) found in the data records. In applications of scan statistics to datasets with fine spatial resolution, such as census tracts or block groups, such highly skewed data distributions are likely to occur. If the spatial distribution estimation process does not handle the zones with low counts correctly, bias in the determination of statistically significant clusters will occur.

In any 8-week baseline period, some of the sparse-data zones have no counts at all. If ignored, the zero-count spatial zones will result in division by zero in the log-likelihood ratio evaluation. The traditional method of setting a floor on the expected counts in each spatial zone leads to a loss of sensitivity when the number of zero count zones is a significant fraction of all the zones. One alternative method for estimating spatial probabilities (APM) is to add one count to the sum of baseline counts in each spatial zone. This method has been used in a study of spatial cluster detection using medical 911 call data from San Diego County with good results [3]. However, when this method was applied to data with a more highly skewed spatial distribution, issues were uncovered which led to this investigation of alternatives.

METHODS

Counts of visits to Texas Department of Defense (DoD) clinics, grouped by syndrome, were acquired for the three year period from 2004 through 2006. The data were spatially resolved into 1233 patient residence zip codes and separately into 32 clinic zip codes, and prospective spatial cluster detection processing was applied for both levels of spatial resolution. For the Poisson spatial scan statistic, a 28-day sliding baseline data window, separated from the test day by a 2-day buffer, was used to estimate representative background spatial probabilities, using a stratified, weekend/weekday (W2) average to handle the marked day-of-week effect evident in this data. Several alternative spatial probability estimation methods were employed and the resulting cluster rates were compared to assess the relative performance of each method. Four syndrome groups were considered; respiratory - with an average of 790 counts per day, GI - with 237 counts per day, rash - with 86 counts per day, and neurological - with 80 counts per day, to examine the impact of count density over an order of magnitude range.

RESULTS

The examined methods for handling the zero-baseline problem include a traditional approach adapted from biostatistics and two alternatives based on improving the expected distribution. For each method, scan statistics were applied to determine significant clusters using both the residence and clinic zip code resolutions for approximately 3 years of daily data. Overall cluster determination rates, frequency of single-zone clusters, goodness-of-fit, and other measures were used for comparison. The presentation will explain the APM 4 method, which gave modest cluster rates and relatively few single-region clusters for both rich and sparse syndrome groups as well as a savings in processing time. For several syndrome groups, goodness-of-fit of this method was as good as that of the space-time permutation adjustment method, but without bias resulting from conditioning on same-day data.

CONCLUSIONS

Alternative methods for the estimation of spatial probabilities and expected counts from counts in a baseline data window are evaluated for prospective cluster detection at fine-resolution with highly skewed spatial distributions having many sparse spatial zones. The relative merits of the methods are assessed for the chosen syndrome groups from the Texas DoD clinic visit counts at the residence-zip code and home-zip code levels of aggregation..

REFERENCES

[1] Kulldorff M (2001) Prospective time-periodic geographical disease surveillance using a scan statistic. J R Stat Soc A Stat Soc 164: 61–72.

[2] Kulldorff M, et al. (2005) A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine, 2:216-224.

[3] Edgerton J, et al. (2006) Space-Time Cluster Detection Using San Diego County 9-1-1 Call Data, PHIN 2006.

Further Information: James Edgerton, James.Edgerton@jhuapl.edu