THE DISTRIBUTION OF INTERPOINT DISTANCES, CLUSTER DETECTION, AND SYNDROMIC SURVEILLANCE

Al Ozonoff, Marco Bonetti, Laura Forsberg, Caroline Jeffery, Marcello Pagano
Harvard School of Public Health, Department of Biostatistics
655 Huntington Avenue, Boston, MA 02115 *

KEY WORDS: Biosurveillance, Clustering, Interpoint distance distribution, Disease mapping, Density estimation

1 Introduction

As concerns rise over terrorism, and in particular an attack via biological or chemical means, biosurveillance has grown in scale as a public health endeavor. To this point, the focus of methodological research intended to support biosurveillance has been almost exclusively aberration detection (some examples from a large collection of literature include [2, 5]). Indeed, the primary purpose of a biosurveillance system is to provide early warning of a potential bioterrorist attack, or perhaps some other public health event that may require further monitoring or response.

Aberration detection has generally followed two parallel tracks. For systems that collect only volume data, an alarm is sounded if there is an excess of disease above what is expected at that point in time [7]. In addition, some systems also have available spatial data, that is the locations of individual cases of disease (either exact locations via geocoding, or perhaps aggregated data at some administrative unit such as census tract or zip code). Detection of spatial or spatio-temporal aberrations comprises a distinct subset of the research literature [2, 3, 5, 8]. There is evidence that if spatial data is available, its utilization can enhance and augment the power of detection of temporal methods [6].

The probability of a bioterrorist attack may be small, which argues for exploring alternate uses of data collected during biosurveillance. There is fewer work available in the literature detailing methodology for such "dual use" efforts. In this paper, we describe a novel method of density estimation that fits within a framework of distance-based aberration detection described in other work [1, 3, 4, 6, 9]. We follow an approach that bears similarities to the image processing techniques of tomography, and apply this technique to surveillance data collected for the purposes of biosurveillance. We propose that such an application can increase understanding of local patterns of disease, thus proving useful regardless of the probability of an event of bioterrorism.

2 Distance-based methods

As noted above, previous work has detailed an approach to analysis of spatial data which can be broadly classified as "distance-based", that is an analysis of spatial patterns based on the (Euclidean) distances between observations or from a fixed point. The M-statistic of Bonetti and Pagano [1] is one example of such a distance-based approach, and this methodology has been successfully incorporated into the context of biosurveillance. The appeal of distance-based methods is the reduction in dimensionality; what can be a difficult problem in two or more dimensions may be more easily solved and well-studied in one dimension (where the bulk of classical univariate statistical theory is available).

We extend the aberration detection capabilities of a surveillance system via density estimation within a distance-based framework. Ordinary density estimators of spatial observations, e.g. kernel density estimators (KDEs), suffer from some limitations, particularly in two dimensions. The underlying (baseline) spatial distribution may be highly heterogeneous, requiring a normalization of the observed density; the number of observations may be small; and the optimal choice for kernel function may be difficult to anticipate. For an application to surveillance data, density estimates should be robust to small sample sizes and

*The first author has a primary appointment at the Boston University School of Public Health, 715 Albany Street, Boston, MA 02118. Data provided courtesy of Institute for Health Metrics (http://www.healthmetrics.org). The research in this paper was funded in part by a grant from the National Institutes of Health, RO1-AI28076, and the National Library of Medicine, RO1-LM007677.

underlying heterogeneity.

Consider a region R in the plane and underlying density $f(x, y)$ which we would like to estimate for all points $(x, y) \in R$. For a given set of observed points $(x_1, y_1) \ldots (x_n, y_n)$, we seek to estimate the observed density normalized by $f(x, y)$. Let C be a circle enclosing R and fix $N > 3$; we set N equally spaced points around the circumference of C. Denote these points by $s_1 \ldots s_N$.

For each point s_i, we construct a one-dimensional density estimate $\gamma_i(d)$ of the observed points $(x_1, y_1) \ldots (x_n, y_n)$ while controlling for the underlying density f as follows. For a given distance d^*, fix a constant k such that $0 < \frac{1}{k} < 1$ and let h be the smallest constant such that the annulus A bounded by radii $(d^* - \frac{h}{2})$ and $(d^* + \frac{h}{2})$ has the property

$$\int_A f d\mu = \frac{1}{k}$$

where $d\mu$ is the ordinary Lebesgue measure on the Euclidean plane. Then the normalized density estimate from s_i at distance d^* is a simple function of the proportion of observed points that fall within the annulus A:

$$\gamma_i(d^*) = \sqrt{n} \cdot \left(\left[\frac{1}{n} \sum_{i=1}^{n} 1((x_i, y_i) \in A) \right] - \frac{1}{k} \right)$$

where $1(\cdot)$ denotes the ordinary indicator function. We rescale by $\sqrt{n}$ since asymptotically $\gamma_i(d) \sim N(0, \sigma^2)$ as $n \to \infty$ under the hypothesis that the observed points are distributed according to f.

Having computed γ_i for each of the points $s_1 \ldots s_N$, we assemble a two-dimensional density estimate Γ by averaging the γ_i for each point $X \in R$ at the appropriate distance:

$$\Gamma(X) = \frac{1}{N} \sum_{i=1}^{N} \gamma_i(d(X, s_i))$$

where $d(X, s_i)$ is the ordinary Euclidean distance from X to each point s_i. As a linear combination of normally distributed random variables, we have $\Gamma(X)$ is normally distributed for a fixed point X with mean zero, under the hypothesis that $(x_1, y_1) \ldots (x_n, y_n) \sim f$.

We note that the method of estimation outlined above is mathematically similar to that of tomographic imaging, where indirect observations (expressed as line integrals of the underlying density) are assembled into an image of the original density via the inverse of the Radon transform. We forego the use of the Radon transform here because we have access to the direct observations. However, our approach shares a mathematical lineage with the (more difficult) problem of tomography.

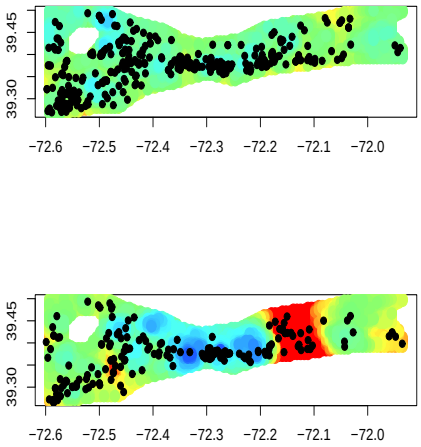

Figure 1: Density estimate adjusted for underlying population heterogeneity on Cape Cod. Shown here are respiratory cases during a flu outbreak (top) and one year later when flu season was not as severe. Red indicates high density (relative to underlying population), dark blue indicates low density.

For practical purposes the level sets of Γ can be converted to a fixed color scale, thus displaying the estimated density across R in sensible colors such as red for high densities and blue for low densities.

3 Application to syndromic data

The method described above was applied to emergency department data from the Cape Cod region of Massachusetts, collected by the Institute for Health Metrics from three Cape Cod hospitals. The data consist of spatial locations of patients arriving for emergency care (as determined by geocoding the patient billing address, coordinates slightly altered to protect anonymity), together with the syndrome group of the patient complaints (using ESSENCE-II syndrome groupings for ICD-9 codes). Data were available on a daily basis for nearly five years between 1994 and 1999.

In order to better understand patterns of disease in the Cape Cod region, we used the estimation procedure described above to map incident disease cases on

a daily basis. We used a 7-day moving average so as to avoid weekly effects. To estimate the underlying density at baseline, we used all cases from one month prior to the days of interest to estimate f. Each (overlapping) 7-day period was then mapped and displayed on a fixed color scale.

Results of the mapping (Figure 1) show that at the height of a major flu outbreak, when respiratory case load was much higher than expected, the spatial distribution of cases follow the baseline spatial distribution closely. In the following year, the flu season was not as severe but the spatial distribution deviated from baseline. Constructing many of these images and viewing them consecutively, as a movie, helps to illustrate the dynamic patterns of respiratory disease as they evolve throughout flu season and at other times of the year.

We conclude that these methods of density estimation while controlling for an underlying heterogeneous population can be effective tools in a syndromic surveillance setting. Further investigation into methods appropriate for syndromic data may yield further insights into the spatial and spatio-temporal patterns of disease in this setting.

References

[1] Bonetti, M and Pagano, M. The interpoint distance distribution as a descriptor of point patterns: An application to cluster detection. *(Accepted, Statistics in Medicine)* (2004).

[2] Burkom, HS. Biosurveillance applying scan statistics with multiple, disparate data sources. *J. Urban Health* **80** (Suppl 1):57-65 (2003).

[3] Olson KL, Bonetti M, Pagano M, Mandl KD. Real time spatial cluster detection using interpoint distances among precise patient locations. *Submitted, BMC Public Health* (2004).

[4] Ozonoff A, Bonetti M, Forsberg L, Pagano M. Revised power comparisons for an improved disease clustering test. *(Accepted, Comp. Stat. Data Anal.)* (2004).

[5] Kulldorff M. A spatial scan statistic. *Commun. Statist. Theory and Methods* **26**: 1481-1496 (1997).

[6] Ozonoff A, Forsberg L, Bonetti M, Pagano M. A bivariate method for spatio-temporal syndromic surveillance. *MMWR* **53** (Suppl):61-66 (2004).

[7] Reis BY, Pagano M, Mandl, KD. Using temporal context to improve biosurveillance. *Proc. Nat. Acad. Science* **100** (4):1961-1965 (2003).

[8] Tango, T. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* **19**: 191-204 (2000).

[9] Whittemore AS, Friend N, Brown BW Jr, Holly EA. A test to detect clusters of disease. *Biometrika* **74**:631-635 (1987).