THE USE OF MULTIPLE ADDRESSES TO ENHANCE CLUSTER DETECTION

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

KEY WORDS: Biosurveillance, Clustering, Interpoint distance distribution, Multiple addresses

1 Introduction

Traditional cluster detection techniques have relied on a number of statistical methods, either to detect deviations from complete spatial randomness or to differentiate between spatial distributions of two or more populations. In most cases the data are presumed to include a set of spatial coordinates, and typically in biostatistical applications the data are coordinates suitably projected into Euclidean space $\mathbb{R}^2$.

We first restrict our attention to the following situation: there is an underlying population distribution, which we will call the *null distribution*. We then consider a subpopulation, perhaps a group of cancer patients, and ask whether the spatial distribution of the subpopulation exhibits spatial deviation (e.g. clustering) from the null. There are a multitude of statistics designed to test the hypothesis that the two distributions differ; examples can be found throughout the literature (e.g. [1, 2, 3, 4, 6]).

Our choice of a test statistic in this setting uses the *distribution of interpoint distances*. The approach is outlined below (Section 2) as well as in [2].

In this paper we consider an extension of these techniques when confronted with a data set that includes two or more spatial locations per observation. In this case, one needs a suitable generalization of the ordinary test statistics that will efficiently utilize the multiple spatial information. The strength of this approach is to allow detection of more complex or subtle potential differences between the null population and the subpopulation of interest.

The multiple address problem can arise in a number of biostatistical settings, for example an environmental study where part of the data collection includes not only residence but also place of employment. We have been investigating the application of this cluster detection methodology to the problem of *biosurveillance*, or routine collection and monitoring of health data in order to detect unusual patterns of disease. In this context, multiple addresses might arise as residential and work address, or some other spatial location such as a school.

We begin this paper with a discussion of the approach of interpoint distances as a method for detecting spatial clustering. The extension of this approach to multiple addresses follows, and we conclude with results of a simulation study and discussion.

2 Distribution of interpoint distances and the M-statistic

We summarize the approach and relevant details of the test statistic M discussed in [2] for distinguishing between the spatial distributions of two populations. By the distribution of interpoint distances we mean the collection of all possible pairwise distances calculated from a collection of points, considered as a distribution curve. For any population distribution, there is a well-defined distribution of distances $\mathcal{F}$ associated with that population. The distance distribution $\widehat{\mathcal{F}}$ calculated from a random sample, drawn from the underlying population, is a consistent estimator of $\mathcal{F}$ in the sense that $\sqrt{n}\,(\widehat{\mathcal{F}}(\cdot)-\mathcal{F}(\cdot))$ converges weakly to a zero mean Gaussian process [2]. An unusual pattern of disease may cause a distortion in the distribution of distances. Measuring the deviation of the observed interpoint distance distribution from that which we expect forms the foundational concept of the M-statistic.

In practice the test statistic is calculated in two steps. First, consider the null distribution by calculating all possible (or a sufficiently large sample of) pairwise distances between individuals in the underlying population. We summarize the distribution by binning the distances into k bins. This yields a k-dimensional vec-

*The research in this paper was funded in part by a grant from the National Institutes of Health, AI28076.

tor describing the distribution of interpoint distances for the underlying population. Repeated calculations of this sort via a resampling or bootstrap procedure provides an estimate of the variance-covariance matrix for the bin counts.

Then for any other subpopulation, we can calculate the pairwise distances between individuals in this group, bin the distances accordingly, and compare the observed distribution to the null. The actual statistic M is a Mahalanobis-like distance: a quadratic form inversely weighted by the variance-covariance matrix. For observed counts o, expected counts e, and estimated covariance matrix S, we calculate:

$$M = (o-e)'\, S^{-}\, (o-e)$$

We can make inferences on the difference between the distribution of the subpopulation as it compares to the null by empirical methods, or in special cases by using the asymptotic distribution of the test statistic (which is known from the theory of U-statistics).

Naturally, the distribution of interpoint distances is highly dependent on the particular metric used. Since most often our data consist of points in Euclidean 2-space, the intuitive metric to use is the standard Euclidean metric on $\mathbb{R}^2$. However there is no *a priori* reason to choose this metric over any other measure of distance or dissimilarity.

3 Multiple addresses

For ease of language, we refer to the situation where a data set contains two or more spatial locations per observation as that of “multiple addresses”. The prototypical example of such a data set is motivated by a syndromic surveillance system under development at Children’s Hospital in Boston by Mandl, et al. [5] Real-time collection of geocoded patient data is analyzed for potential clusters of disease outbreak. The residential location of each patient is part of the data set. The research group is attempting to incorporate the school that each patient attends (or an equivalent second address such as a work address) into the data set, and develop analytic methods for analyzing such data.

The most straightforward analysis of multiple address data would involve testing each address independently, and applying some correction to the alpha level of each test to account for the multiple testing problem, for example a Bonferonni adjustment. Although this approach allows for more than one model of clustering (i.e. clustering in any of the multiple addresses), the multiple testing correction may greatly reduce power. Indeed, because the addresses are (possibly highly) correlated, the test statistics on each address are not independent. Finally, this approach will have greatest power to detect a cluster where spatial clustering occurs solely in one particular address. There may be reduced power in situations where the pattern of clustering is more complex across a mixture of addresses.

Another approach is to consider a pair of observations, and to choose an appropriate metric on $\mathbb{R}^2 \times \cdots \times \mathbb{R}^2$ and calculate the interpoint distance between these two observations. The M-statistic could then be calculated as before. This offers the advantage of reducing the dimensionality of the problem to one where we can apply the standard univariate methods. However without any advance knowledge of the underlying pattern of disease, the particular method chosen to measure distance may not be especially sensitive to a potential cluster. Worse still, there might be a significant penalty associated with specifying a metric that is inappropriate for certain situations.

To circumvent the necessity of choosing such a metric, we instead propose to consider the distribution of distances in each address jointly. A binning procedure to aggregate the joint distance distribution into a single vector (defined over $\mathbb{R}$) then allows us to proceed as before with the calculation of the M-statistic.

4 Simulations

Our simulations were designed to assess the performance of the 2-dimensional M-statistic as it compares to the 1-dimensional (ordinary) M. We use a simplistic model to generate data and simulate disease outbreaks. Consider the unit circle as a study area, with a large population of individuals whose residential addresses are distributed uniformly. Across the study area are a number of “schools” (what we think of as our second address), also distributed uniformly. Each individual is assigned to a school according to a pre-specified probability function that is conditional on the proximity of the school location to the individual location. With an appropriate choice of this probability function, we can exhibit a high degree of spatial correlation between the population’s home and school addresses.

Figure 1 illustrates the type of data simulated. The smaller points indicate residential addresses in the population. A small sample from three sub-populations (corresponding to individuals attending each of three different schools) has been color-coded and plotted. The figure illustrates the high degree of correlation between residential and school address.

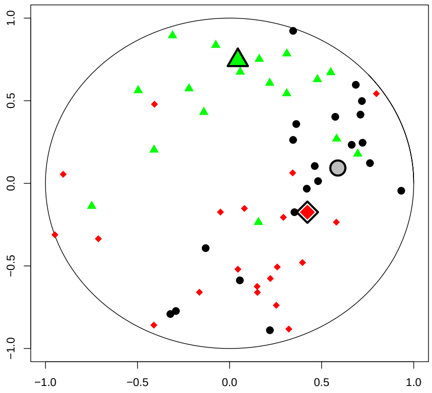

Figure 1: Simulated data with both home and school addresses. Smaller points are individuals; large outlined points are school locations. Addresses exhibit a high degree of correlation.

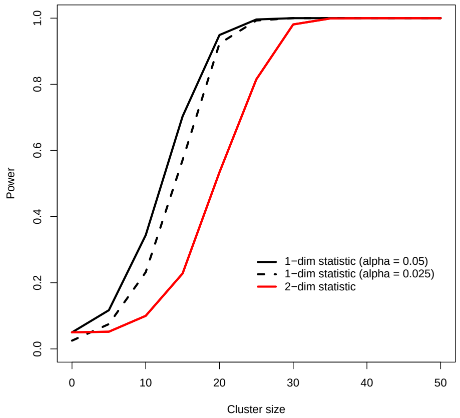

Figure 2: Power to detect a local cluster, for the 1- and 2-dimensional M-statistics. The 2-dimensional statistic shows a slight loss of sensitivity.

After establishing the null distribution of the M-statistic (as described above) in both the 1-dimensional and 2-dimensional versions, we examine the power to detect different types of clusters. The difference in power is meant to demonstrate the benefits or drawbacks of utilizing more than one address in cluster detection.

We would expect the 1-dimensional M-statistic that utilizes only the home address to outperform the 2-dimensional M when detecting local clusters. The difference in the power to detect these clusters may be seen as the penalty one must pay for incorporating an additional address when the pattern of disease follows the home address.

To simulate a cluster, we first took a random sample of 300 individuals as background case load. Then, a variable number of these individuals were replaced with "cluster cases" that were selected according to some criteria depending on the type of cluster. For example, in one scenario all of the cluster cases attend the same school irrespective of the home address; we call this a "school cluster". Likewise, if all of the cluster cases are in one localized (residential) location irrespective of school address, we call this a "local cluster".

Simulations varied the cluster size from 5 cases to 50 cases, superimposed on the background case distribution for a total of 300 cases. We estimated the power using 1000 simulated clusters, in both a local cluster and school cluster model, using the 1-dimensional M-statistic and the 2-dimensional M. Since the Bonferroni adjustment is conservative, we calculated power for the 1-dimensional M using an alpha level of both $.05$ and $.025$, and the 2-dimensional M with an alpha level of $.05$.

Results of the power calculations are summarized in the Figures 2 and 3. Figure 2 shows the power curves for the 1- and 2-dimensional statistics as they performed on local clusters; Figure 3 shows the same curves for school clusters.

5 Discussion

Figure 2 makes clear that there is a cost in power to detect when utilizing the second address for local clusters. We expect the 1-dimensional statistic to outperform in this scenario, and it does so even with a Bonferonni adjustment. The greatest difference is seen when the signal strength is weakest; as the cluster size grows, the 2-dimensional statistic detects clusters with high power despite its lower sensitivity.

Figure 3 demonstrates the benefit in considering both addresses, since the 2-dimensional statistic outperforms when the clustering model is in the school address. From the standpoint of the 1-dimensional statistic this is a misspecified alternative, thus the power to detect is quite low for the 1-dimensional statistic despite the high correlation in addresses. Since the addresses are correlated we might use one address as a proxy for the other; this provides evidence that the loss

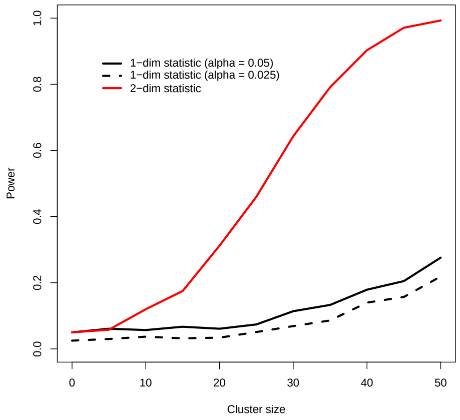

Figure 3: Power to detect a school cluster, for the 1- and 2-dimensional M-statistics. The 1-dimensional statistic suffers from a misspecified alternative.

of power is significant in this case.

Many questions have been left unexplored in this simulation study, for example the gains and losses in sensitivity of each statistic as we vary parameters in the simulation model. Those parameters include the underlying population distribution (which is taken to be homogeneous here for simplicity); the degree of correlation between addresses; and the effects of missing data in one or both addresses. All of these issues seem highly relevant to a working implementation of a multiple address system. There are certainly other candidates for a statistic that incorporates multiple addresses, and the performance of these candidates should be investigated as well.

The availability and utility of more than one address in a biosurveillance setting will depend on the particulars of the situation. Based on these and other ongoing simulations, the evidence suggests that in cases where a system must be robust to more than one pattern of clustering, utilization of all available information is important to maintaining an adequate level of detection. On the other hand if multiple address collection places a heavy burden on the system, or if there is no reason to believe that there is more than one model for disease outbreak, maintaining the highest power for this single alternative seems wise and one address only is probably most appropriate. Nonetheless, further research into the methodology and characteristics of multiple address systems should provide a valuable foundation for more sophisticated future surveillance systems.

References

[1] Alexander, F and Boyle, P. [eds.] *Methods for Investigating Localized Clustering of Disease*. IARC Scientific Publications No 135 (1996).

[2] Bonetti, M and Pagano, M. The interpoint distance distribution as a descriptor of point patterns: An application to cluster detection. *(Submitted, Statistics in Medicine)*, 2003.

[3] Cressie, NAC. *Statistics for spatial data*. Wiley-Interscience (1991).

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

[5] Mandl, K and Lee TH. Integrating medical informatics and health services research: the need for dual training at the clinical health systems and policy levels. *J. Am. Med. Inform. Assoc.* 2002; **9**: 127-132.

[6] Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 1967, 209–220.