Early Detection of Tuberculosis Outbreaks among the San Francisco Homeless: Trade-offs Between Spatial Resolution and Temporal Scale

Higgs BW[†], Mohtashemi M[†‡], Grinsdale J[*], Kawamura LM[*]

[†]MITRE Corporation, 7515 Colshire Dr., McLean, VA, 22102 USA; [‡]MIT Department of Computer Science, 77 Mass. Ave., Cambridge, MA 02139-4307 USA; [*]TB Control Section, San Francisco Dept. of Public Health, Ward 94, 1001 Potrero Ave., San Francisco, CA 94110 USA

Background

San Francisco has the highest rate of tuberculosis (TB) in the US. Exact locations of patients' primary residences at the time of diagnoses are routinely collected as part of the TB surveillance program. It has been shown for syndromic surveillance data that when exact geographic coordinates of individual patients are used, higher detection rates and accuracy are achieved compared to when data are aggregated into administrative regions such as zip codes and census tracts. Here, we examine the effect of varying the spatial resolution in the TB data on the San Francisco homeless population, on detection sensitivity, timeliness, and the amount of historical data needed to achieve better performance measures.

Methods and Findings

We apply a variation of space-time permutation scan statistic to the TB data in which a patient location is either represented by its exact longitude and latitude or by the centroid of its census tract. We show that the detection sensitivity and timeliness of the method generally improve when exact locations are used to identify both simulated and real TB outbreaks, however, better performance measures were attained under simulated cases as compared to actual outbreaks. Finally, we compare the dependency of the method on the extent of data needed for parameter estimations under different geospatial constraints, and show that smaller amount of data is required when exact locations are used to achieve similar performance measures.

Conclusion

We investigate the relationship between using exact locations of TB patients, the timeliness of identifying real TB outbreaks, and the amount of legacy data required for early detection. We demonstrate that using higher spatial resolution results in higher detection rate, but more importantly in timely detection of TB outbreaks even when the amount of available data is relatively small. For higher spatial resolution, we also show generally better sensitivity for simulated outbreaks as compared to actual outbreaks, though this difference can be explained by the variations in dispersal structure of cases between the two. Trading higher spatial resolution for better performance, however, is ultimately a tradeoff between maintaining patient confidentiality and improving public health. Understanding such tradeoffs is critical to managing the complex interplay between public policy and public health. This study is a step forward in this direction.

INTRODUCTION

TB is one of the top four diseases for infection-induced mortality in the world today. There are currently about 54 million people infected with the bacterium *Mycobacterium tuberculosis* with approximately 8 million new infections occurring each year. TB kills nearly 2.4 million people annually. In the U.S. alone, there are currently about 12.5 million people who have been infected by TB (Ginsberg 2000), with the city of San Francisco having the highest rate in the U.S. Although in recent years the incidence of TB has been declining in the San Francisco general population (see Figure 1), it has remained relatively constant in the homeless population (see Figure 2).

Spatial investigations of disease outbreaks seek to identify and determine the significance of spatially localized disease clusters by partitioning the underlying geographic region. The level of such regional partitioning can vary depending on the available geospatial data on cases including towns, counties, zip codes, census tracts, and exact longitude-latitude coordinates. When exact patients' locations have been used in cancer surveillance, the accuracy was not appreciably higher than that obtained with larger and more conventional regional partitions such as census block groups (Gregorio et al., 2005), though the benefit of localized rate variation (i.e. geographic excess or shortfalls in cancer incidence for small areas) was shown in an earlier study (Gregorio et al., 2002). In other works, there were few performance differences observed for larger aggregation comparisons such as block group, census tract, zip code, and town (Krieger et al, 2002; Sheehan et al., 2000). More recently, Olson, et al, using the method of space scan statistic (Kuldorff et. al., 1997) applied to syndromic data, has shown that when patients' exact locations are used, higher detection rates are achieved as compared to center points of larger geographical regions such as zip codes and census tracts (Olson et al., 2006). The authors demonstrated that the advantage in using higher resolution cluster detection is primarily in a reduction of the distortion effect that is induced by the use of large detection windows (i.e. spatial scanning windows), as compared to smaller detection windows. This problem occurs when, for example, two cases are geographically close to one another, while they reside in two separate zip codes (or census tracts). In such situations, if the geographic partitioning is by zip code (or census tract), the detection window has to be rather large to encompass both administrative regions since the cases are represented by the centroids of these regions, while a smaller detection window can capture such localized cases when the exact individual addresses are used.

In space-time surveillance of disease outbreaks, however, the interdependency between both time and space are manifested by disease clusters that are localized in time and space. Such disease localizations can be investigated through dynamic partitioning of the underlying geographic regions, where different degrees of spatial resolution can be coupled with varying levels of temporal scale to examine both the detection rate and timeliness. While the benefits of using higher spatial resolutions, such as patients' individual addresses, have been examined in the context of spatial epidemiology, the spatio-temporal effects of disease localizations have not been studied under different degrees of spatial aggregation. As such, the effect of varying degrees of spatial aggregation on detection timeliness has not been investigated. At the same time, any detection method must rely on

a pool of legacy data to both establish a baseline of normal disease variability and estimate the model parameters. However, the amount of available data varies across surveillance programs, and historical data are often in short supply. Therefore, in addition to detection sensitivity and timeliness, the dependency on the amount of historical data must also be examined when varying spatial resolution. Finally, under the multiple levels of aggregation, the bias in geographic spread, that can be introduced when creating simulated cases can cause differences when comparing simulated cases to the spread of actual outbreaks. This disparity between simulated and actual cases must also be examined to make meaningful analogies for progressing from validation in a synthetic environment to validation in a real surveillance system.

In this work, we use a modification of space-time permutation scan statistic (Naus 1965; Kulldorff 1997; Kulldorff et al., 2005; Wallenstein 1980; Weinstock 1981) to examine the effect of varying degrees of spatial resolution (census tracts of patients' residences versus exact locations) on the detection sensitivity and timeliness using both simulated and confirmed outbreaks applied to the TB data on the homeless population of San Francisco for 1991-2002. We find that when exact patient's locations are used, the detection method can identify more outbreaks (designated or confirmed) and, more importantly, in a more timely manner for both simulated and actual cases. Finally, we show that with individual addresses the detection method requires a smaller amount of historical data to achieve similar performance measures obtained under census tract centroids.

MATERIALS and METHODS

Data

The San Francisco Department of Public Health (SFDPH), TB Control Program (TBCP), routinely collects comprehensive information on TB cases and their contacts including demographic (e.g., age, gender, race), population risk factors (e.g., intravenous drug use, HIV status, alcohol intake), laboratory results (e.g., skin test, chest x-ray), time of diagnosis, and primary residences. For the homeless population, the geospatial information typically includes shelters and single room occupancies (SROs). In addition to the above data types, advances in molecular biology have made it possible to identify different bacterium fingerprints with the technique of restriction fragment length polymorphism (RFLPs) and polymorphic GC-rich repetitive-sequence (PGRS) methodologies (Small et al., 1994; van Deutekom et al., 1997). With these technologies, it is possible to both identify and track specific subpopulations that have been infected with the same bacterial strain. This information can aid in outbreak investigation to identify patterns and hubs of transmission often hidden in a network of complex interactions between primary infected cases and their contacts (Klovdahl et al., 2001; McElroy et al., 2003).

The dataset for this study consists of comprehensive information on 392 individuals that have been diagnosed by the SFDPH, TBCP with active TB and identified as homeless over the time period of 1991-2002. The primary residences of these individuals were used to identify their geographical coordinates (latitudes and longitudes) using ArcGIS v9.0

(ESRI). The census tract information for identifying the tracts in which the homeless individuals reside, were obtained from generalized extracts from the Census Bureau's TIGER geographic database provided by the US Census Bureau (http://www.census.gov/geo/www/cob/index.html). There were a total of 76 unique census tracts covered by the study population. TB case data is kept electronically in a patient management database maintained by the SFDPH, TBCP. All case information, including address of residence and homeless status at the time of diagnosis, was downloaded directly from the database. Census tract information was obtained from the 2000 census.

Confirmed Outbreaks— p9 cluster

During 1991-2002, an epidemic strain of TB took hold among the homeless population in San Francisco. Both RFLP and PGRS analyses were conducted on infected cases to identify the particular strain and associate it with previously identified molecularly similar clusters. This investigation resulted in 47 unique homeless individuals being identified as infected carriers of this new strain not previously observed, and referred to as the *p9 cluster*. This cluster arose at two separate time periods, peaking in 1996 and disappearing by 1999, with a second outbreak rapidly appearing in 2001 (how about a plot here?).

Modified space-time permutation scan statistic

Variations to both the scan statistic introduced by Kulldorff et al., 2005 and the method for fast detection of spatial overdensities, provided by Neill et al., 2003, is implemented here as a suitable method for space-time investigation of TB outbreaks in the San Francisco homeless population. For a more detailed account of the method, see the work cited above.

Briefly, the method can be described as follows. Instead of using circles of multiple radii as spatial bases for scanning cylinders (Kulldorff et al., 2005), a square grid approach, similar to that provided by (Neill et al., 2003) is employed here. Overlapping grids containing p squares, each of area r^2 are placed over the entire region, where the grid overlap is permitted at half the width of each square, representing the spatial domain. The time domain, as in (Kulldorff et al., 2005), is represented by the length of such (rectangular) cylinders. For each square, the expected number of cases, conditioned on the observed marginals is denoted by μ where μ is defined as the summation of expected number of cases in a cylinder, given by $\mu = \sum_{(s,t)\in A} \mu_{st}$ where s is the spatial cluster and t is the time span used (e.g., days, weeks, months, etc.) and $\mu_{st} = \frac{1}{N}(\sum_{s} n_{st})(\sum_{t} n_{st})$, where N is the total number of cases and n_{st} is the number of cases in either the space or time window (according to the summation term). The observed number of cases for the same cylinder is denoted by n. Then the Poisson generalized likelihood ratio (GLR), which is used as a measure for a potential outbreak in the current cylinder, is given by

$\left(\frac{n}{\mu}\right)^{n}\left(\frac{N-n}{N-\mu}\right)^{(N-n)}$ (Kleinman et al., 2005). To assign a degree of significance to the GLR value for each cylinder, Monte Carlo hypothesis testing (Dwass 1957) is conducted, where the observed cases are randomly shuffled proportional to the population over time and space and the GLR value is calculated for each square. This process of randomly shuffling is conducted over 999 trials and the random GLR values are ranked. A p-value for the original GLR is then assigned by where in the ranking of random GLR values it occurs.

For our space window, we restricted the diameter of squares to range from 0.02 km to 1 km. For different sized squares that had a perfect intersect of the same cases, the smallest square was retained. There were a total of 441 space scanning squares sampled for the census tract centroids, and a total of 4,234 for individual residences. For our time window, the TB case count is much lower than the daily data feeds typical of surveillance systems used to monitor emergency room visits due to the influenza-like illnesses or pharmacy sales, for example. To compensate for the smaller proportion of total cases, daily counts were aggregated into monthly (approximately 4 weeks) case counts resulting in a total of 144 data points (months). This agglomeration of cases is necessary since the notion of *early detection* of outbreaks of a chronic disease, such as TB, with long incubation period (Benenson 1995) requires a longer time scale. This is in contrast with sydromic surveillance of acute infectious disease (such as influenza-like illnesses) where early detection encompass only hours to days after the start of an outbreak (see Mandl et al., 2004; Lewis et al., 2002; Reis et al., 2003; Mohtashemi et al., 2006; Mohtashemi et al., 2007). Finally, the amount of historical data used for training the model and parameter estimations varied from 4-72 weeks, spanning the years of 1991-2002.

Reduction of Overlapping Signals

Due to the number of likelihood ratios calculated, multiple testing correction becomes an important procedure for determination of significant signals. The Monte Carlo hypothesis testing step is designed to correct for much of this. However, it is often the case that multiple signals with similar significance also share a high degree of similar information, such that a procedure for reduction of such redundant information is required. For example, a geographical square deemed significant that has two cases within the four week window may intersect with another square, also deemed as significant for the same time window, that is larger in size, contains three cases, and encompasses the first square. Under such circumstances when there is a 100% intersect of a smaller square with a larger square for the exact same time period and similar significance measure, a unique signal is reported by retaining the smaller square. This procedure was performed on all significant signals using the above criteria

RESULTS

Simulated Outbreaks—Detection sensitivity and timeliness

We examine the detection power and timeliness of the application of the space-time detection method previously described to TB data infused with simulated outbreaks under both individual addresses and census tract centroids. Similar to the approach in (Olson

et al., 2005) the scanning window size was increased from the smallest region of 0.02 km to the largest region of 1 km in size, while the simulated points were distributed in multiple administrative regions (census tracts) ranging from 1 to 4 (see Figure 3 and Table 1). All simulated cases were first randomly placed in one administrative region, then split into two administrative regions, and so on up to four total regions. When the scanning window is small , the method attains greater detection sensitivity using individual addresses regardless of the number of regions within which the cases are added (see Figure 3a-b and Table 1). As the scanning window size is set to 0.2 km (Figure 3c) and increased to larger values (Figure 3c-e), this trend continues only for the cases distributed over the multiple administrative regions (i.e. 3 and 4). The census tract centroids have greater detection sensitivity when cases are distributed over fewer administrative regions (1 and 2) using scanning window sizes of 0.2 km and greater. At the largest scanning window size (1 km), there are very few cases detected over both individual addresses and census tract centroids, due to the convergence of the cluster grid and overall detection region.

The speed of detection is one of the most important performance measures for disease surveillance. In the space-time permutation scan statistic, the time window is varied at time spans ranging from 1 month to 6 months. Outbreaks that occur in a short time interval for a specific geographic region or adjacent regions are detected within the small time window size, whereas, those outbreaks that have more sparse case counts with time (i.e. spread out over a longer time span than 1-2 months) for a specific geographic region or adjacent regions, are detected within a larger time window size. In the context of a disease such as TB with chronic characteristics, we denote an outbreak is detected early if it is identified within months (less than a year) from the start of the outbreak. To assess the differential detection timeliness of the method using individual addresses and census tract centroids, additional cases were simulated at the same frequency and under the same conditions, and over the same geographic regions, while the scanning time window was increased from 1-6 months. As can be seen in Figure 4, at increasing time windows greater than 2 months, more significant clusters are detected with the use of individual addresses. At time windows less than 3 months, no noticeable difference is observed between the two methods. Both individual addresses and census tract centroids demonstrate a linear relationship with time after 2 months.

Confirmed Outbreaks—Detection sensitivity and timeliness

Here, we examine the detection sensitivity and timeliness of the method using the two levels of geographical resolutions (i.e. census tract centroids and individual addresses) applied to the confirmed outbreaks of the p9 cluster (Table 2 and Figure 5). To provide parity between the two methods, some assumptions were made in the case where both approaches were detecting the same signal, with overlapping, yet slightly different start and end dates. If the start or end dates for a significant signal under one geographic resolution overlapped the dates for the other, we considered them the same signal, That is, the detection method identified the same signal under both geographic constraints. The initial increase in cases that occurred in 1995, as well as the continuation of this outbreak through 1996, 1997, and the large resurgence in 2002 were detected under both spatial resolutins (see Table 2). These p9 clusters are accurately detected at three

separate time points (1995-1996, 1997, and 2002) that are consistent with the documented outbreaks of this epidemic strain. It should be noted that the two outbreaks detected in 2002 using individual addresses overlap the single outbreak detected in 2002 using census tract centroids, so this is treated as a single detection by each method.

With exception to the significant cluster that is detected within the dates of 11/26/96-12/17/96 using individual addresses (increased sensitivity), there is not a considerable difference in the detection sensitivity of the two. Though one could argue that this additional cluster that was detected using individual addresses in 1996 (and not detected by census tract centroids) is evidence that supports overall better performance with individual addresses (as compared to census tract centroids), based on the list of 4 principle time points with confirmed p9 outbreaks. By assessing accuracy as a ratio of the number of detected (and confirmed) outbreaks to the total number of confirmed outbreaks, the use of individual addresses detects 100% of the confirmed outbreaks, whereas the use of census tract centroids detects 75% of the confirmed outbreaks. Irrespective of this last point,,the detection timeliness is improved when using individual addresses. Here, the timeliness is assessed by the detection time window, which constitutes the number of weeks that are used within the base of the scanning grid to calculate the expected and observed number of cases (used for the generalized likelihood ratio). When examining the time window lengths for similar signals under the two spatial resolutions, the detection method using individual addresses generally requires smaller time windows than that of census tract centroids. As can be seen from Table 2, the significant signal detected within the dates of 5/13/97-8/26/97 using individual addresses is detected approximately 1 month prior the same outbreak is detected using census tract centroids. In addition, the end date of this time window, using individual addresses, is approximately 2 months earlier. This earlier end date pattern is also demonstrated for the 2002 outbreak as well, even when the use of individual addresses identifies a significant cluster with two separate date blocks. For detection of identical outbreaks, the use of individual addresses (as compared to census tract centroids) is advantageous in reporting the significant cluster in a shorter time interval.

Legacy data requirement

The accuracy of surveillance algorithms is often determined by assessments on retrospective data. Documented outbreaks in the past are used as ground truth to measure model performance. Depending on the algorithm, the larger the pool of legacy data, the more reliable are the performance measures, since the assessment of normal variability improves with measurements over time We assessed the accuracy of the algorithm using the two spatial resolutions by simulating increasing cases on a background of real TB cases, using varying amounts of historical data. First, we randomly selected a geographic region, and simulated 1-8 cases over the range of 4-72 weeks (roughly 1-18 month) of historical data. The percentage of simulated cases with the background of real cases in the scan window was then recorded, as well as the detection p-value from the Monte Carlo hypothesis test for that scan window. For example, if 3 cases are simulated within a background of 3 cases, this region has had a 100% (3/3) increase in cases, 50% (3/6) consisting of simulated cases, and the detection p-value calculated may be 0.20. Now if 3 more simulated cases are added to the previous 3 simulated cases and the background of

3 cases, this region now has had a 200% (6/3) increase in cases, 67% (6/9) consisting of simulated cases, and the detection p-value calculated may be 0.01. We implemented this methodology using the p-values (Figures 6a and 6b y-axis) and simulated case percentages (Figures 6a and 6b: x-axis), starting with only one month of legacy data to draw upon, all the way up to 18 months. Then this procedure was repeated 1,000 times for different randomly selected regions.

Figures 6a and 6b illustrate the sensitivity of significant signals to the amount of historical data required to detect outbreaks within a one-month window. The orange dashed lines represent a significant weak and strong signal and correspond to p-values of 0.001 and 0.0001, respectively. When the availability of historical data is limited to only 2-3 months, individual addresses provide a more sensitive measure than census tracts. With the use of census tracts, there is neither a weak nor strong significant signal detected with 2 months of legacy data, as opposed to the weak signal observed for individual addresses. In addition, with 3 months of legacy data, the use of individual addresses demonstrate detection of a weak signal when simulated cases were added at 90%, whereas, the use of census tracts allows detection of the weak signal, requiring simulated cases to be added at 95%. When the availability of historical data is limited to 6 months, the method detects a weak signal at 90% of simulated cases using census tracts, while it detects a strong signal at about the same 90% using individual addresses. Note that increased detection sensitivity is implied when the detection method identifies fewer simulated cases using smaller amount of legacy data under individual addresses.

DISCUSSION

We investigated the effect of varying the spatial resolution in a variant of a widely used space-time detection technique on the sensitivity and timeliness of identifying both simulated and confirmed TB outbreaks, and examined the dependency of these performance measures on the amount of historical data required. We showed that when exact patients' locations are used, irrespective of whether the outbreaks were simulated or real, both performance measures are generally improved compared to when census tracts are used as the spatial base for geographic partitioning. Furthermore, using simulated outbreaks we demonstrated that when individual addresses are used, the detection method requires smaller amount of historical data than that required to achieve similar performance measures under census tract centroids. Overall, higher performance improvements were achieved under simulated outbreaks, compared to real TB outbreaks, when individual patients' coordinates were used for search.

The results of Table 1 warrant some discussion. First, when individual patients' coordinates are used as basis for geographic partitioning, the detection method consistently performs better if the size of the scanning window is sufficiently small (<0.2 km). This appears to be true regardless of the degree of case-spread in the study region (see Table 1). Second, when the size of the spatial scanner is sufficiently large (0.2 km or larger) to encompass an entire area covered by one or two census tracts, if the cases are clustered within such an area, then the method performs better under census tract-based partitioning. Third, when the simulated cases were spread over more than two

administrative regions, further expansion of the spatial base did not result in better performance using census tracts, yet the detection sensitivity was better using individual coordinates under similar conditions (see Table 1).

When simulating cases for the sensitivity comparison, the cases were randomly distributed within 1-4 administrative regions (census tracts), such that the spread of cases was approximately uniform. This was conducted in such a manner as to not bias the spread of cases around any particular region. We attempted to measure the approximate case-spread under both the simulated and real outbreaks using two measures—the coefficient of variation (CV) in the pair-wise distance distribution of simulated cases and the CV of the distribution of distance between cases and center of their census tracts. The first measure gives an indication of the extent of spread of the simulated cases with respect to each other while the second is an indication of how spread these cases are with respect to the center of their census tracts. The respective CVs were 0.85 and 1.88 and the respective variances were 0.39 and 0.6. In principle, the larger the variance, or the CV, of a distribution, the closer is the underlying distribution to that of uniform, the distribution by which these cases were simulated and the results of Table 1 were obtained. This means that such spatially uniformly distributed cases are much more spread with respect to center of their administrative regions than from each other. In such a situation, expanding the spatial base of the search when cases are spread over more than two regions does not improve the detection. On the other hand, the detection method performed better using individual addresses even when the spatial base was large and cases were spread over four census tracts.

For the confirmed outbreaks, the respective CVs for the pair-wise and case-to-census distance distributions were 0.69 and 1.33 while their variance was the same (0.2), which can partially explain why the improvement in detection sensitivity using individual locations was not as considerable as with the simulated cases. However, using exact patients' coordinates, the detection method almost invariably was able to identify localized clusters of smaller sizes earlier in time, which is a critical property of real time surveillance and timely containment of disease outbreaks.

While the results of this study clearly point to improvement in the detection sensitivity and timeliness when patients' coordinates are used as the center of the spatial scanner, the larger improvement was obtained when the cases were randomly generated. This is a sensible result, because the real TB case distribution in the event of an actual outbreak is not expected to characterize a uniform distribution. Factors affecting the spread and transmission of TB in the homeless population, such as localization of shelters and SROs to specific geographic regions and the high prevalence of intravenous drug use and HIV and AIDS among the homeless, result in spatial clusters of TB that are topologically different from those attained under random distribution of cases in space. Thus, we infer that quantitating the extent and topology of disease case-spread derived from historical data, can widely benefit real time surveillance and guide public health investigations with respect to detection and control of infectious diseases. While the decision on which spatial resolution results in improved detection sensitivity may depend on localization properties of historical case spread, we showed that the detection timeliness is

consistently improved when the detection method uses patients' coordinates as the center of its spatial base for search.

Finally, trading higher spatial resolution for increased performance is ultimately a tradeoff between maintaining patient confidentiality and improving public health. While these features are critical to real time surveillance, maintaining patient confidentiality introduces a challenge to the timely investigation of outbreaks. The complex interplay between public policy and public health may be better managed by understanding and balancing the associated risks in each problem domain. As critical as this topic of debate may be, it is outside the scope of this work.

REFERENCES

Bailey, TC. and Gatrell, AC. 1995 Interactive Spatial Data Analysis, Second Edition: Longman

Benenson AS (ed). Control of Communicable Diseases Manual. American Public Health Association, Washington , DC, 1995.

Dwass M. (1957) Modified randomization tests for non-parametric hypotheses. *Ann Math Statist*, **29**:181187.

Ginsberg, A. (2000) A Proposed National Strategy for Tuberculosis Vaccine Development. *Clinical Infectious Diseases,* **30**:S233-242.

Gregorio DI, Kulldorff M, Barry L, Samociuk H. (2002) Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, Connecticut, 1991-95. *Int J Cancer,* 10;100(2):194-8.

Gregorio DI, Dechello LM, Somciuk H, Kuldorff M. (2005) Lumping or splitting: seeking the preferred areal unit for health geography studies. *Int J Health Geogr*., 4:6.

Hirschfield, A., Yarwood, D. and Bowers, K. (1997) Crime Pattern Analysis, Spatial Targeting and GIS: The development of new approaches for use in evaluating Community Safety initiatives, N. Evans-Mudie (ed) Crime and health data analysis using GIS, Sheffield: SCGISA.

Kleinman KP, Abrams AM, Kulldorff M, Platt R (2005) A model-adjusted space-time scan statistic with application to syndromic surveillance. *Epidemiol. Infect*., 000:1-11.

Klovdahl AS, Graviss EA, Yaganehdoost A, Ross MW, Wanger A, Adams GJ, Musser JM. (2001) Networks and tuberculosis: an undetected community outbreak involving public places. *Soc Sci Med*., **52**(5):681-694.

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. (2002) Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? *Am J Epidemiol*, 156:471-482.

Kulldorff M. (1997) A spatial scan statistic. *Commun Stat A Theory Methods*, **26**:1481-1496.

Kulldorff M, Heffernan R, Hartmann J, Assuncao R, Mostashari F (2005) A space-time permutation scan statistic for disease outbreak detection. *PLOS*, 2(3).

Lewis MD, Pavlin JA, Mansfield JL, O'Brian S, Boomsma LG *et al.*, (2002) Disease outbreak detection system using syndromic data in the greater Washington, DC area, *Am. J. Prevent. Med.* **23** (2002) (3), pp. 180–186.

Mandl KD, Overhage JM, Wagner MM, Lober WB, and Sebastiani P *et al.*, (2004) Implementing syndromic surveillance: a practical guide informed by the early experience, *J. Am. Med. Inf. Assoc.* **11** (2004) (2), pp. 141–150.

McElroy PD, Rothenberg RB, Varghese R, Woodruff R, Minns GO, Muth SQ, Lambert LA, Ridzon R. (2003) A network-informed approach to investigating a tuberculosis outbreak: implications for enhancing contact investigations. *Int J Tuberc Lung Dis*.S486-93.

Mohtashemi M, Szolovits P, Dunyak J, Mandl KD (2006) A susceptible-infected model of early detection of respiratory infection outbreaks on a background of influenza. *J Theor Biol*, 241(4):954-63.

Mohtashemi M, Kleinman K, Yih K (2007) Multi-syndrome analysis of time series using PCA: A new concept for outbreak investigation. To appear in *Stat Med.*

Naus J. (1965) The disctribution of the size of maximum cluster of points on the line. *J Am Stat Assoc*, **60**:532-538.

Neill DB, Moore AM. (2003) A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems*, 16

Olson KL, Grannis SJ, Mandl KD. (2006) Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health*, 96(11):2002-2008.

Openshaw, S. (1984) The modifiable areal unit problem. *Concepts and Techniques in Modern Geography* 38:41.

Reis BY, Pagano M and Mandl KD, Using temporal context to improve biosurveillance, *Proc. Natl. Acad. Sci. USA* **100** (2003), pp. 1961–1965.

Sheehan TJ, Gershman ST, MacDougal L, Danley RA, Mroszczyk M, Sorensen AM, Kulldorff M. (2000) Geographic surveillance of breast cancer screening by tracts, towns and zip codes. *J Public Health Manag Pract*, 6:48-57.

Small P M, Hopewell P C, Singh S P, Paz A, Parsonnet J, Ruston D C, Schecter G F, Daley C L, Schoolnik G K. (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med*, **330**:1703–1709

Unwin, DJ. (1996) GIS, spatial analysis and spatial statistics. *Progress in Human Geography* 20(4):540-441.

van Deutekom H, Gerritsen J J J, van Soolingen D, van Amijden E J C, van Embden J D A. (1997) A molecular epidemiological approach to studying the transmission of tuberculosis in Amsterdam. *Clin Infect Dis*, **25**:1071–1077

van Emdben JA, Cave MD, Crawford JT, et al. (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*. 31, 406-409.

Wallenstein S. (1980) A test for detection of clustering over time. *Am J Epidemiol*, **111**:367-372.

Weinstock MA. (1982) A generalized scan statistic test for the detection of clusters. *Int J Epidemiol*, **10**:289-293.

FIGURES

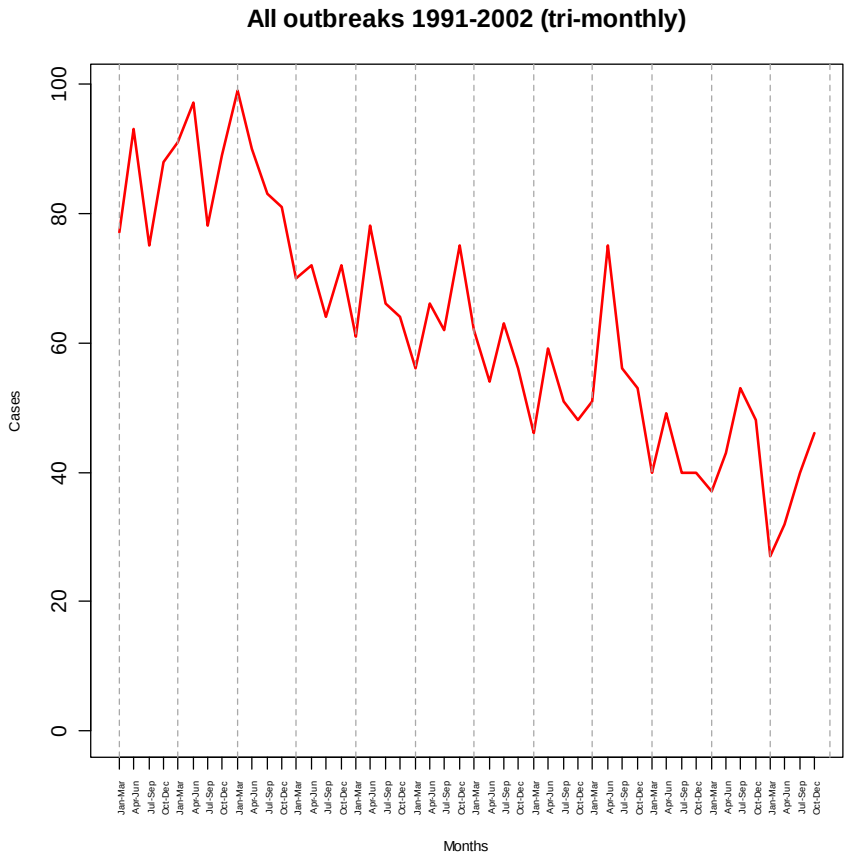

Figure 1. Three month interval temporal plot of TB cases in the San Francisco general population for the years of 1991-2002. Grey dashed lines separate each year.

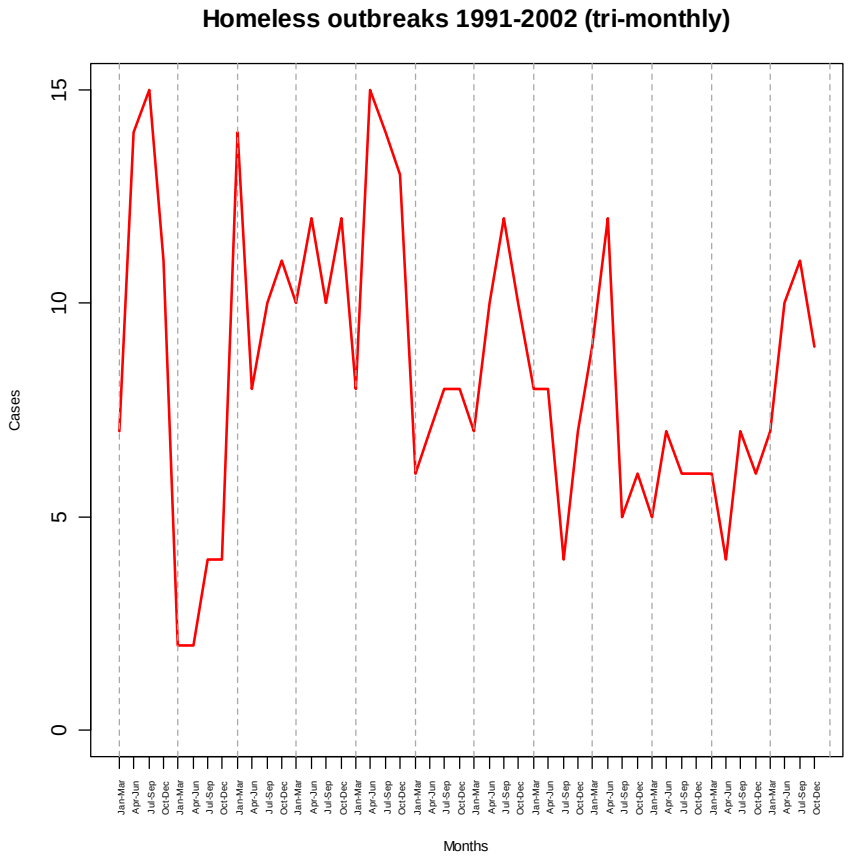

Figure 2. Three month interval temporal plot of TB cases in the San Francisco homeless population for the years of 1991-2002. Grey dashed lines separate each year.

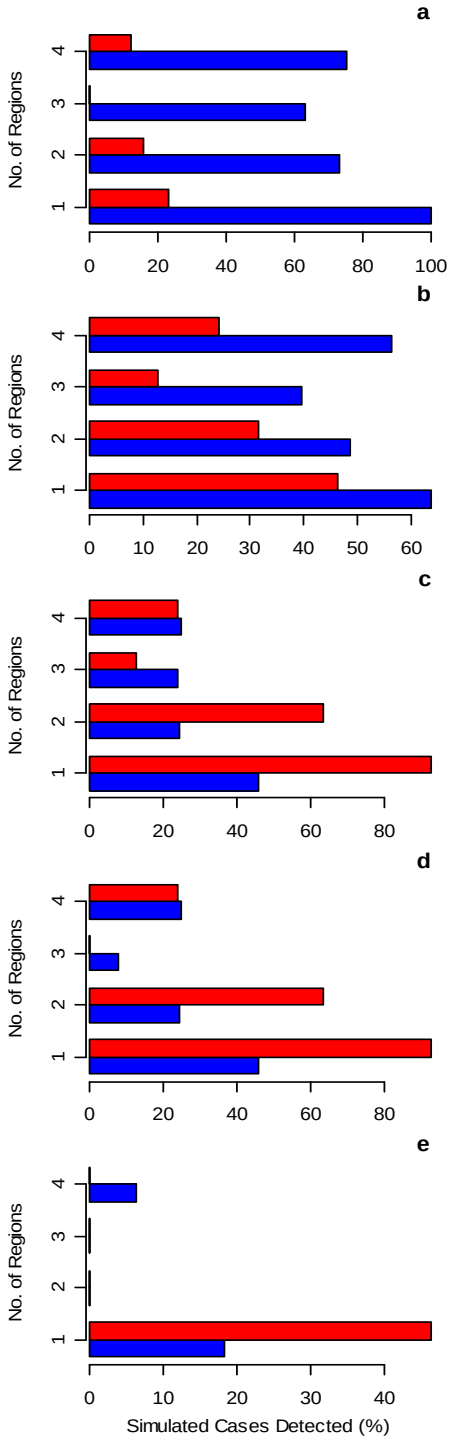

Figure 3. Detection sensitivity of simulated cases. Significant clusters detected for individual addresses (blue bars) and census tract centroids (red bars) with increasing spatial windows of a) 0.02 km, b) 0.1 km, c) 0.2 km, d) 0.5 km, e) 1 km, and census tract regions (1 to 4) . For each plot, the x-axis is scaled to the maximum detection sensitivity percentage.

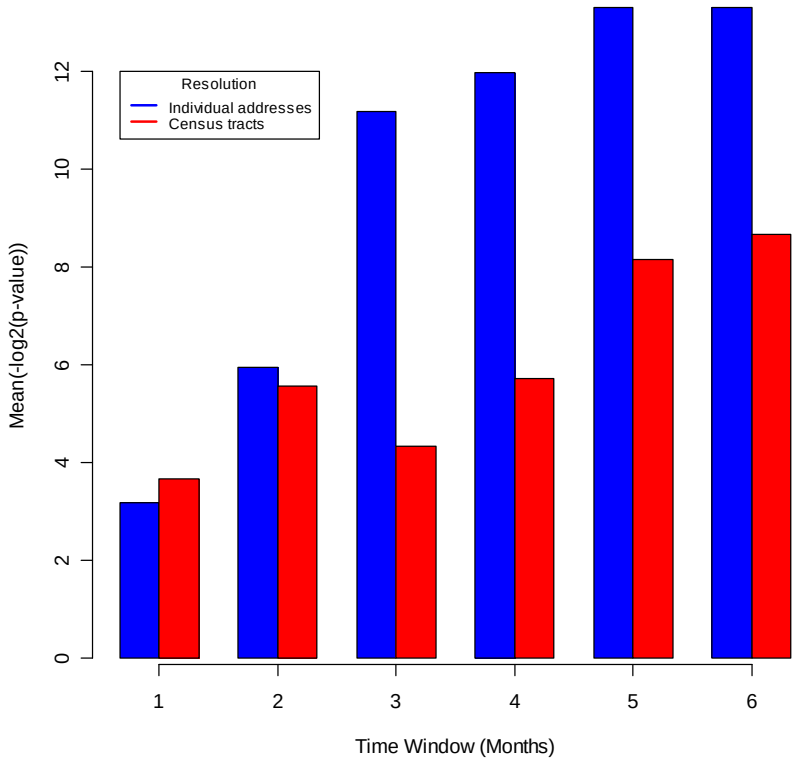

Figure 4. Detection timeliness of simulated cases. The average -log2 transformed p-value distribution for individual addresses (blue bars) versus census tract centroids (red bars) with an increasing window size.

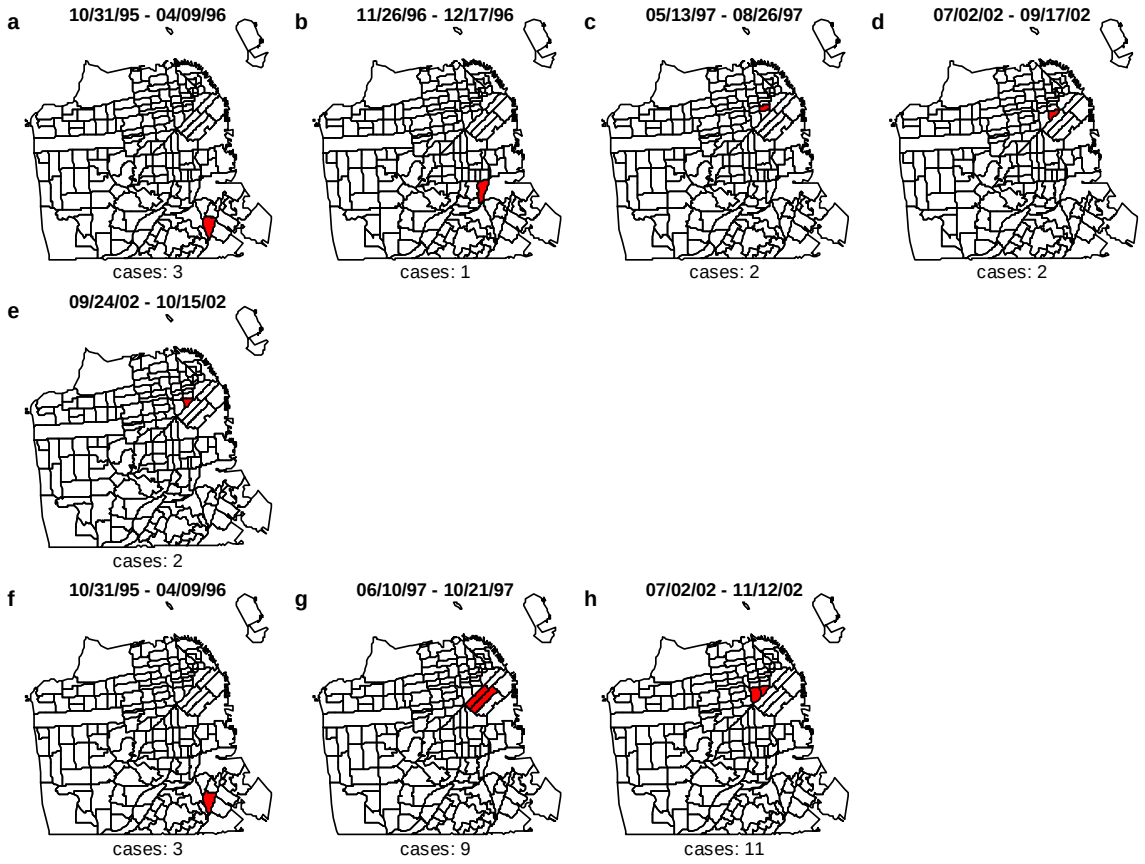

Figure 5. Significant signals for the p9 outbreaks using individual addresses (a-e) and census tract centroids (f-h). Bay area maps partitioned by census tract where tracts shaded red represent the location of the significant signal detected for the specified dates in Table 2. To protect patient confidentiality, only the census tracts in plots a-e are shaded, as opposed to highlighting the exact locations.

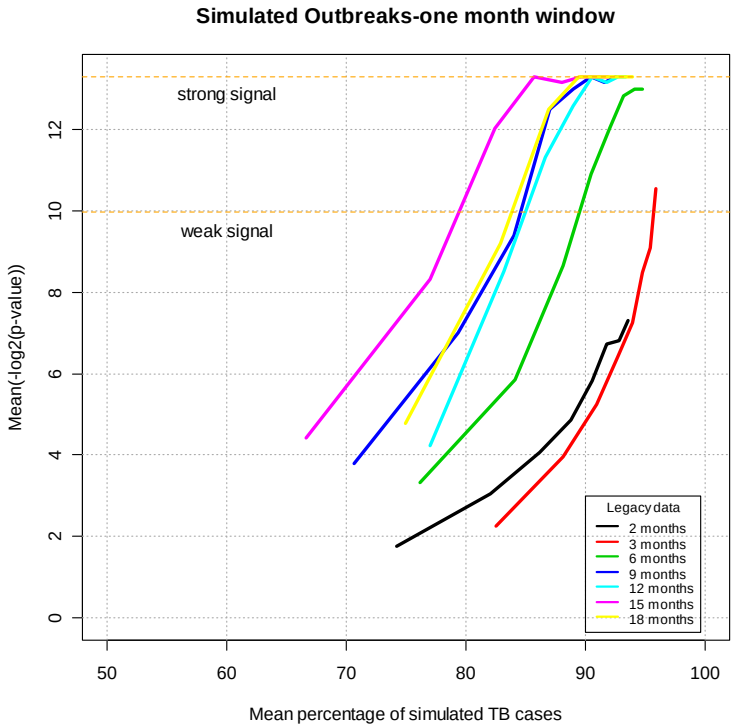

Figure 6a. Sensitivity of detection to historical data. Significance of detected simulated outbreaks to the amount of legacy data required using census tract centroids. The orange dashed lines represent a significant weak and strong signal correspond to p-values of 0.001 and 0.0001, respectively.

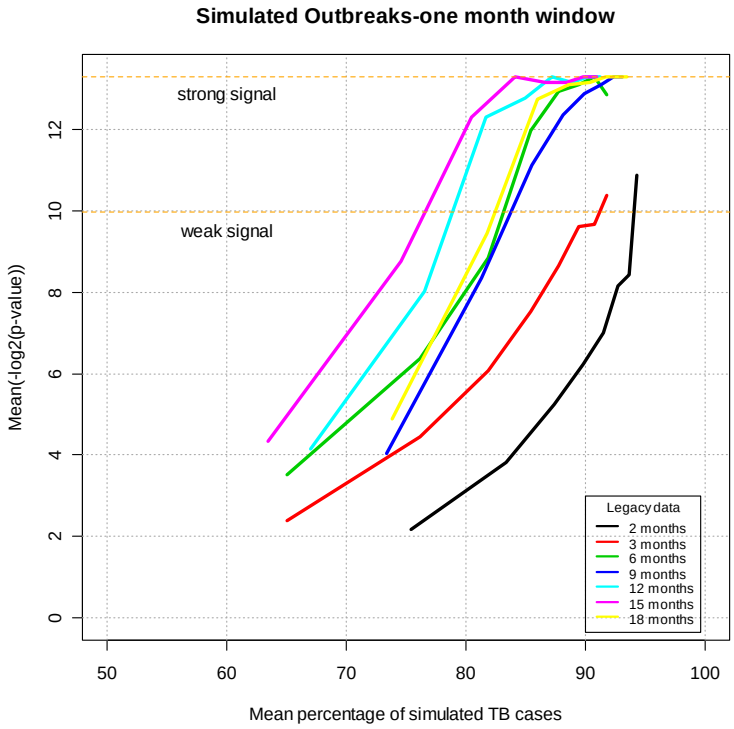

Figure 6b. Sensitivity of detection to historical data. Significance of detected simulated outbreaks to the amount of legacy data required using individual addresses. The orange dashed lines represent a significant weak and strong signal correspond to p-values of 0.001 and 0.0001, respectively.

TABLES

Table 1. Significant clusters detected for individual addresses and census tract centroids with increasing spatial windows from 0.02 km to 1 km. Data in table corresponds to Figure 3.

0.02 km	**individual address**	**census tract**
1	100.00%	23.12%
2	72.97%	15.92%
3	63.36%	0.00%
4	75.08%	12.01%

0.1 km		
1	63.66%	46.25%
2	48.65%	31.53%
3	39.64%	12.91%
4	56.46%	24.02%

0.2 km		
1	45.65%	92.49%
2	24.32%	63.36%
3	23.72%	12.91%
4	24.92%	24.02%

0.5 km		
1	45.65%	92.49%
2	24.32%	63.36%
3	7.81%	0.00%
4	24.92%	24.02%

1 km		
1	18.32%	46.25%
2	0.00%	0.00%
3	0.00%	0.00%
4	6.31%	0.00%

Cells shaded grey represent regions where individual addresses have a higher detection percentage than census tract centroids.

Table 2. Detection timeliness and number of significant p9 clusters using census tract centroids and individual addresses.

	Individual addresses		**Census tract centroids**	
Outbreak	Start date	End date	Start date	End date
1	10/31/1995	4/9/1996	10/31/1995	4/9/1996
2	11/26/1996	12/17/1996		
3	5/13/1997	8/26/1997	6/10/1997	10/21/1997
4	7/2/2002	9/17/2002	7/2/2002	11/12/2002
	9/24/2002	10/15/2002		