Approved for Public Release; Distribution Unlimited
Case # 06-1211

A SPACE-TIME SCAN STATISTIC FOR DETECTION OF TB OUTBREAKS IN THE SAN FRANCISCO HOMELESS POPULATION

BRANDON W. HIGGS[†]

MITRE Corporation, 7515 Colshire Dr.
McLean, VA, 22102 USA

MOJDEH MOHTASHEMI

MITRE Corporation, 7515 Colshire Dr.
McLean, VA, 22102 USA

MIT Department of Computer Science, 77 Mass. Ave.
Cambridge, MA 02139-4307 USA

JENNIFER GRINSDALE

TB Control Section, San Francisco Dept. of Public Health, Ward 94, 1001 Potrero Ave.
San Francisco, CA 94110 USA

L. MASAE KAWAMURA

TB Control Section, San Francisco Dept. of Public Health, Ward 94, 1001 Potrero Ave.
San Francisco, CA 94110 USA

San Francisco (SF) has the highest rate of TB in the US. Although in recent years the incidence of TB has been declining in the general population, it appears relatively constant in the homeless population. In this paper, we present a spatio-temporal outbreak detection technique applied to the time series and geospatial data obtained from extensive contact and laboratory investigation on TB cases in the SF homeless population. We examine the sensitivity of this algorithm to spatial resolution using zip codes and census tracts, and demonstrate the effectiveness of it by identifying outbreaks that are localized in time and space but otherwise cannot be detected using temporal detection alone.

[†] contact: bhiggs@mitre.org

1. Introduction

Tuberculosis (TB) is one of the top four diseases for infection-induce 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).

Though advances in health and medicine had considerably reduced the incidence of TB after the “mid” 20th century, there was an increase in cases in many parts of the United States in the mid-1980s, and early-1990s, in part due to increased homeless individuals and prevalence of AIDS (which compromises the immune response). In San Francisco, annual TB cases peaked in 1992 with “51.2 cases per 100,000 persons and decreased significantly thereafter to 29.8 cases per 100,000 persons in 1997” (Jasmer et al., 1999). Currently TB case rates are approximately 20 per 100,000 persons. An important characteristic of TB is that once droplet nuclei containing bacteria are expelled into the air, such droplets are able to circulate in confined spaces so that direct and prolonged contact with an infectious person is no longer a prerequisite to acquire infection (Wells 1934; Wells et al., 1934). This factor is of particular importance in the homeless population, where individuals typically seek care at locations of limited space, such as shelters or single room occupancies (SRO)s, which can amplify a single exposure quickly.

With such an efficient mode of transmission in the homeless population, strategies to mitigate the spread of TB are important. Surveillance systems should be equipped to target the spatial spread within and between shelters and SROs over time to reduce the likelihood of potential outbreaks. Combining this information across different dimensions, such as time and space (Klovdahl et al., 2001), significantly enhances our understanding of the underlying transmission dynamics, and hence will improve upon existing public health intervention policies for control and prevention of TB.

In this paper, we propose the scan statistic first examined in 1965 (Naus 1965) and later implemented in other work (Kleinman et al., 2005; Kulldorff 1997; Kulldorff et al., 2005; Wallenstein 1980; Weinstock 1981) for detection of potential TB outbreaks in the homeless population of San Francisco from the years of 1991-2002. We demonstrate that the scan statistic is a sensitive measure for identifying aberrant frequencies (from normal trends) of TB cases within certain time and spatial distributions that would otherwise go undetected using deviations from a global average or methods that depend only on temporal patterns.

We find that the distribution of TB cases within zip code regions follows a power law distribution, where many individuals with TB reside in a select few zip codes and few individuals with TB reside in many different zip codes. We show that a more resolved clustering of region into census tracts can reduce this skewness in the spatial distribution of TB cases and improve detection sensitivity.

2. Methods

2.1. *San Francisco Department of Public Health TB Data*

TB case data is kept electronically in a patient management database maintained by the San Francisco Department of Public Health, TB Control Section. 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.

2.2. *Space-Time Permutation Scan Statistic Applied to TB Data*

A variation to the scan statistic introduced by Kulldorff (Kulldorff et al., 2005) was implemented here as a suitable method for early detection of TB outbreaks in the San Francisco homeless population, particularly for those time/region-specific increases in case frequency that are too subtle to detect with temporal data alone. Similar to the scan statistic proposed by Kulldorff et al, the scanning window utilizes multiple overlapping cylinders, each composed of both a space and time block, where time blocks are continuous windows (i.e. not intermittent) and space blocks are geographic-encompassing circles of varying radii. Briefly explained here (see Kulldorff et al., 2005 for a more in depth account of the algorithm), for each cylinder, 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

$$\mu = \sum_{(s,t)\in A} \mu_{st} \tag{1}$$

where s is the spatial cluster and t is the time span used (e.g., days, weeks, months, etc.) and

$$\mu_{st} = \frac{1}{N}(\sum_{s} n_{st})(\sum_{t} n_{st}) \tag{2}$$

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

$$\left(\frac{n}{\mu}\right)^{n}\left(\frac{N-n}{N-\mu}\right)^{(N-n)} \quad \text{(Kleinman et al., 2005).} \qquad (3)$$

Since the observed counts are in the numerator of the ratio, large values of the GLR signify a potential outbreak. 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 cylinder. 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 geographic circle to three radius sizes of small, medium, and large: 0.23 miles (0.37 km), 0.44 miles (0.70 km), and 0.56 miles, (0.90 km), respectively. 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 or pharmacy sales, for example. To compensate for the smaller proportion of total cases, monthly case counts were used (with a time window of 2 months) spanning the years of 1991-2002. We observed that the homeless population that we surveyed is a closer approximation to a uniform population at risk as compared to the general population and has less dependence on certain social patterns. For example, unlike associations between emergency department visits and specific days of the week, or medicine sales promotions and increases in sales of medication targeting a specific ailment (Kulldorff et al., 2005), the TB-infected homeless population in not affected by these variables and exhibits greater uniformity in case counts with time. Many of the confounding factors that can influence these case count fluctuations in the general population such as socio-economic status, days of the week, holidays, and season, do not affect the true case counts of TB. We conducted the scan statistic across all 144 months as well as stratifying across seasons, to adjust for the largest confounding variable, but did not observe a large difference in the top scoring GLR space/time combinations between the two methods.

2.3. *Data*

The dataset consists of 392 individuals that have been diagnosed by the San Francisco Department of Public Health with active TB and identified as homeless over the time period of 1991-2002. The primary residences of these individuals have also been identified, where a residence is defined as either a shelter, a single room occupancy (SRO), or a county prison. The total number of zip codes that account for these residences include 22, with 80 census tracts.

3. Results

3.1. *Scale-free property of TB cases in the homeless population (1991-2002)*

When examining the frequencies of TB cases across the 12 year time period within each geographically partitioned region (zip codes) in the San Francisco homeless population, there are a small number of hub zip codes. These hubs are defined as regions with a large density of TB cases as compared to the other regions. Some of these hubs tend to fluctuate in density (i.e., become smaller/larger and less/more dense) in certain years, while others persist as regions of highly dense case counts throughout the 12 year time period, where highly dense is defined as an apparent increase. A viable assumption that explains the existence of these highly dense hubs is the relationship between region population and number of TB cases. One would assume that the more populated regions would have more TB cases, simply as a function of a larger population size. However, when utilizing the 2000 population census data for each zip code, the correlation between these two variables is very low (r<0.15), so adjusting for population (i.e. per capita statistic) does not alter the observed hub regions.

More interesting than the observation of these hub regions is the degree of self-organization that exists, resembling large-scale properties of complex networks (Barabasi et al., 1999). When examining the total number of TB cases in the homeless population for each region across the 12 year time period, the number of cases and regions follow a power law for zip codes (Figure 1 blue points). Similar to large networks that exhibit a power-law distribution for self-organizing into a scale-free state, the average number of TB cases across regions demonstrate the same characteristic of this "universal architecture" (Keller 2005). That is to say, as new connecting edges (k), are added to a node, the edges attach with a probability $P(k)$ proportional to the number of edges already connected to the current node. This decay rate is represented by $P(k) \sim k^{-\gamma}$ where γ is the exponent constant. The rate at which a node acquires edges is given by $\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}$ which gives $k_i(t) = m\left(\frac{t}{t_i}\right)^{0.5}$, where

t_i is the time at which node i was added to the system and m is the number of nodes. So the probability that a node i has a connectivity smaller than k, $P[k_i(t) < k]$ can be written in terms of m as $P(t_i > \frac{m^2 t}{k^2})$ or $1 - P(t_i \le \frac{m^2 t}{k^2})$, which is equal to $1 - \frac{m^2 t}{k^2}(t + m_0)$ (Barabasi et al., 1999). The number of cases (c) is analogous to the edge degree or k, and the cumulative number of regions with c cases is analogous to $P(k)$. The degree of decay for the plot of the cumulative number of regions with c cases versus c represents the distribution property, or how evenly distributed the number of cases are over the regions. A function that decays quickly has a more even distribution of cases over the regions than a function that decays slower. From Figure 1, it is apparent that the distribution of cases over the zip codes is distributed much less evenly than the distribution of cases over census tracts.

Based on this skewed distribution of the hub zip codes, a simple surveillance system might target such TB outbreak “hotspots” for mitigation with the assumption that the primary mode of transmission can be greatly reduced through moderation of the most TB case-rich zip codes. However, as illustrated in Figure 2, the zip codes where individuals with TB reside are more spatially indistinct than clear. There is an averaging effect for each zip code when summating the total number of cases for each, where individuals with TB that reside on the border of two or more zip codes can create a hub that is not identified by a single zip code. For example, within the space between zip codes 1, 2, and 3, there is a large density of residences for TB cases, however, attempts at targeting (for purposes of mitigation) each zip code separately would be inappropriate since the primary density of cases occur within the space between the three zip codes. A scan statistic can address this issue with the addition of more cylinders with finer overlapping space windows. However, the space window will always have to include at least two zip codes to include the density between the regions. To account for this problem and to work with spatial clusters that have TB cases more evenly distributed throughout, we found it more useful to use the smaller units to partition the space, such as census tract.

3.2. *Analysis of homeless population outbreaks (1991-2002)*

Table 1 lists the homeless TB case occurrences for the most significantly scoring space-time windows. Out of the 12 year time period analyzed, specific spatial regions in the years of 1991, 1992, 1993, and 1997 produce significant signals for a high count of TB cases, relative to normal variability. This result of potential TB outbreaks within these years is consistent with previous reports that have documented the early-1990s as the time period where there was a resurgence of TB cases in the San Francisco area (Jasmer et al., 1999).

The top GLR values (most significant signals) correspond to different successive combinations of the months in the range of March to June in 1991 for a small region of census tracts that is contained by all three different sized circles (Figure 3). All three of these circles intersect at some point to share similar information, which explains why the same region exhibits a high signal in multiple time windows. For the regions defined by these 3 circles of increasing radii that do not intersect, one can observe how far the signal extends. When examining only the temporal plot of TB case frequencies in the homeless population, stratified by month for the years 1991-2002 (Figure 4), there are apparent peaks observed in 1991 that correspond to the scan statistic significantly scoring time windows. For example, within the range of March 1991 to October 1991, the number of TB cases in the homeless population never drops below 3 cases. In fact, in 4/12 months for this year there were at least 5 cases of TB. For this example, the frequency peaks are large (compared to normal trends across the 12 year period), such that the temporal information alone could potentially be used to identify the peak signal months in the year 1991, once the cases were mapped to the general residential area.

However, for the other significantly scoring space-time windows in the table within the years of 1992, 1993, 1997, the temporal information alone does not intuitively infer a potential signal. In fact, for some of the significantly scoring space-time windows, the temporal information can be counterintuitive. For example, if one examines the temporal information alone for the year 1992, there are 4 months where 2 cases of TB and 1 month where 1 case of TB is documented (Figure 4). This year accounts for a small number of cases for any one month, such that if one relies only on the temporal information for detection of a significant signal, none of the months in this year would raise a flag. So what information causes this time window for the 3 month combinations of July 1992 to September 1992 to appear as a significant signal? The answer is in the number of TB cases for a particular census tract relative to the total number of cases over the entire 12 year period (Figure 5). That is to say, for the 2 TB cases that are spatially mapped within the same cylinder (1992/smallest circle), there are only 8 total TB cases within this same region over the entire 12 year period. So, 25% of the total TB cases in this region over the 12 year time span occurred in 1992, and as a result, produce a potential signal.

All of the significant signals are plotted in Figure 6 with the dates of case detection labeled. The numbers in the plot correspond to the unique signals denoted in the 'Circle' column of Table 1, to better illustrate the spatial trend of significant frequencies of TB cases over the time period. It is interesting to note the spatial distribution of the significant TB case counts for the 4 years determined. Through the months of March to June in 1991, the significant signals occur in the north central region (represented by circles labeled 1), then from August to September of the same year, the signal predominates slightly northwest (represented by circles labeled 2 and 3), but in the same general northern region. Then in the months of July to September of 1992, the signal occurs back in the northern central region (represented by circle labeled 4), as it had originated in the months of May to June in 1991. In the months of February to April in 1993, the significant signal occurs in a completely separate region in the south central area (represented by circle labeled 5). This is interesting since the signal seems to deviate from the other years in a completely separate region for this year. Finally, in the months of September and October in 1997, the significant signal occurs back near the north central region (represented by circle labeled 6) as had originally occurred for significant signals in the years of 1991 and 1992.

Table 1. Scan statistic results for most significant GLR values.

GLR	p-value	Expected	Observed	Radius*	Circle†	Date1	Date2
629.716	p<0.0001	0.32	4	s,m	1	Mar-91	Apr-91
85831.34	p<0.0001	0.36	6	s,m,l	1	Apr-91	May-91
2868.833	p<0.0001	0.21	4	s,m	1	May-91	
155956.7	p<0.0001	0.32	6	s,m,l	1	May-91	Jun-91
3010.488	p<0.0001	1.28	8	l	2	Aug-91	Sep-91
143.755	p<0.0001	1.18	6	s,l	3	Aug-91	Sep-91
77.675	p<0.0001	0.09	2	s	4	Jul-92	Aug-92
297.477	p<0.0001	0.04	2	s	4	Aug-92	
297.477	p<0.0001	0.04	2	s	4	Aug-92	Sep-92
297.477	p<0.0001	0.04	2	m	5	Feb-93	Mar-93
749.225	p<0.0001	0.03	2	m	5	Mar-93	Apr-93
376.38	p<0.0001	0.64	5	m	6	Sep-97	Oct-97

* Size of space component of cylinder used to group regions: small (s), medium (m), and large (l)

† Unique circle surrounding the regions of significant signal

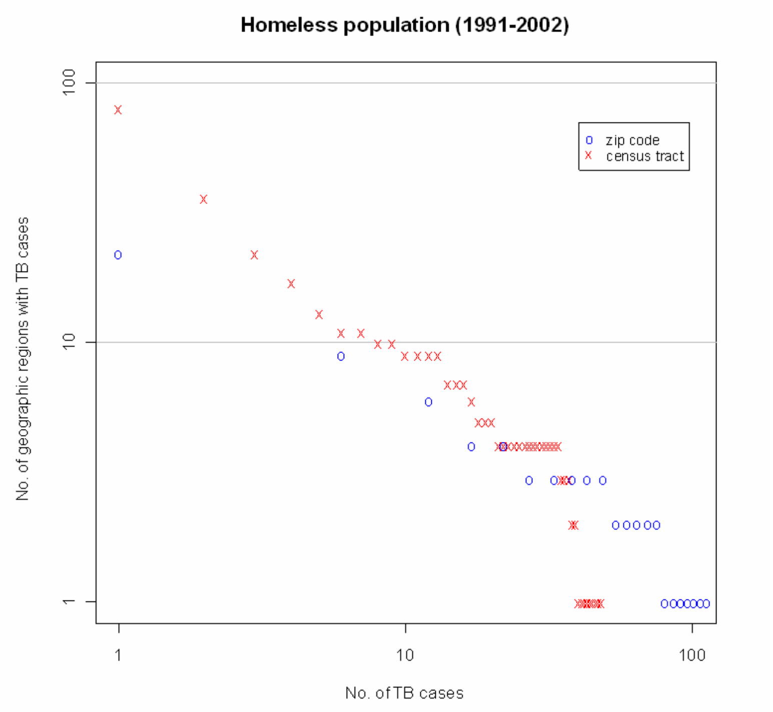

Figure 1. Cumulative distribution of TB cases in the homeless population within both zip codes (blue) and census tracts (red) in the San Francisco area.

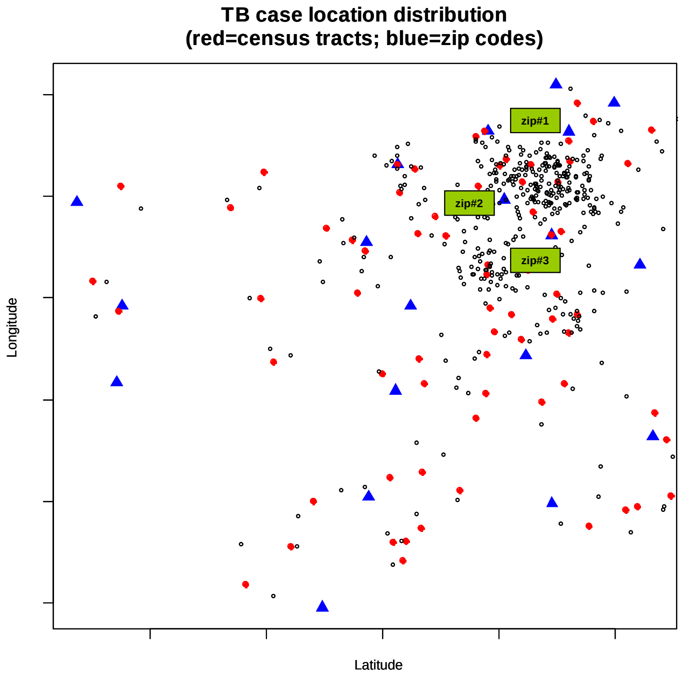

Figure 2. Distribution of TB cases in the homeless population (1991-2002). Black points represent TB cases, red points represent census tracts, and blue triangles represent zip codes. Dense regions of TB cases are not evenly accounted for with zip code partitioning as illustrated in the region between zip codes 1-3.

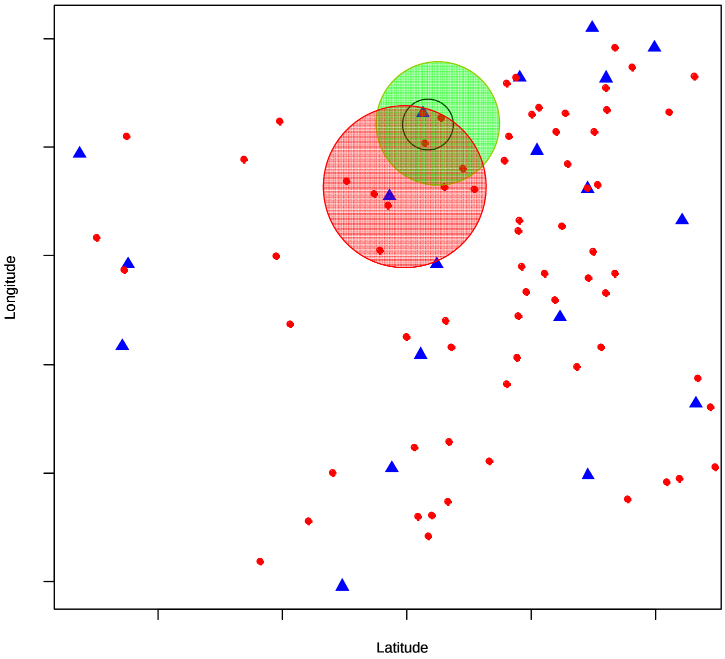

Figure 3. Identified region of significant alarm in 1991. Red points represent census tracts and blue triangles represent zip codes. Three overlapping circle sizes (small: clear, medium: green, and large: red) represent the spatial window for different time windows including the months of March to June in 1991.

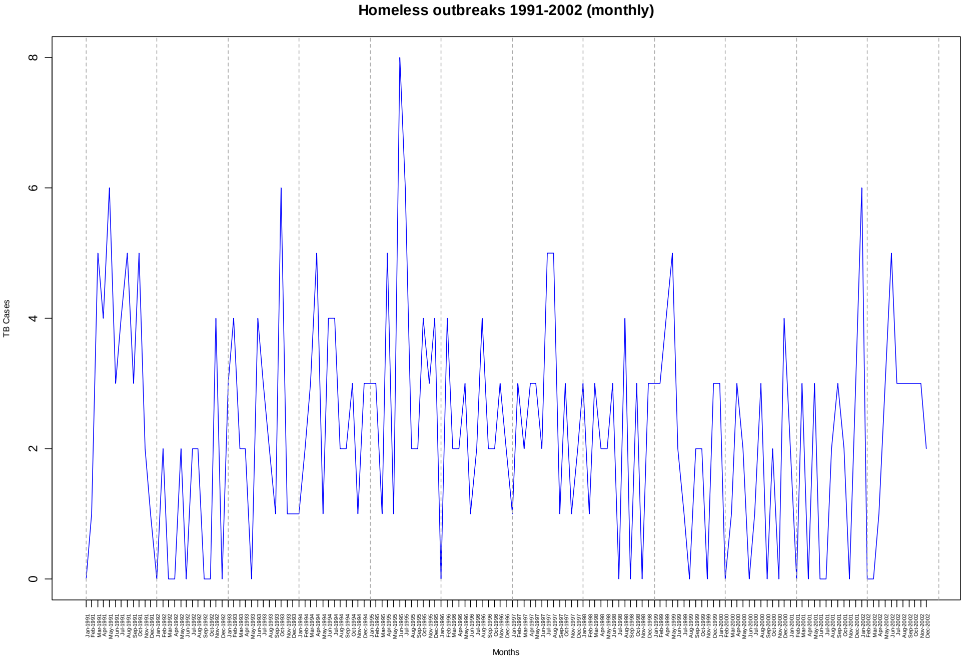

Figure 4. Temporal plot of TB cases in the homeless population for the years of 1991-2002. Data is stratified by months, where the dashed vertical lines divide years.

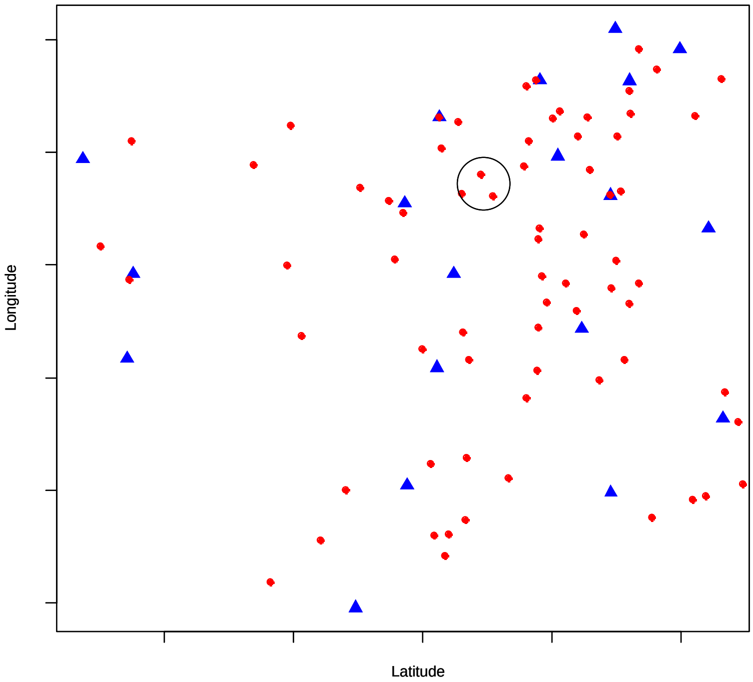

Figure 5. Identified region of significant alarm in 1992. Red points represent census tracts and blue triangles represent zip codes. Small circle size represents the spatial window for different time windows spanning the months of July to September in 1992.

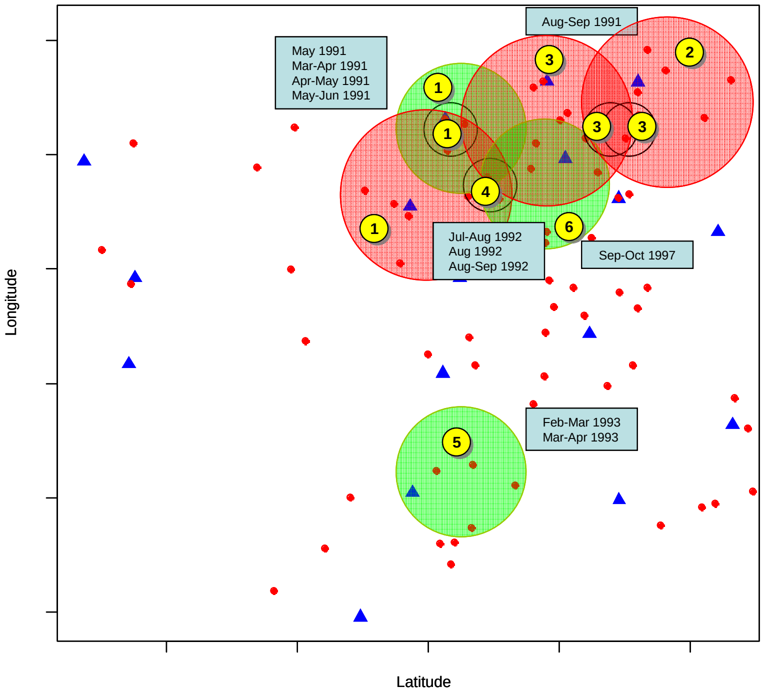

Figure 6. Spatial plot of top scoring signals from Table 1 with different sized circles representing spatial windows (small: clear, medium: green, large: red). Dates for time windows are provided next to each space window.

References

Barabasi AL, Albert R. (1999) Emergence of Scaling in Random Networks. *Science*, **286**:509-512.

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.

Jasmer RM, Hahn JA, Small PM, Daley CL, Behr MA, Moss AR, Creasman JM, Schecter GF, Paz EA, Hopewell PC. (1999) A Molecular Epidemiologic Analysis of Tuberculosis Trends in San Francisco, 1991–1997. *Ann Intern Med*, **130**:971-978.

Keller EF. (2005) Revisiting “scale-free” networks. *BioEssays*, **27**:1060-1068.

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.

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).

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

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.

Wells WF. (1934) On airborne infection. Study II. Droplets and droplet nuclei. *Am J Hyg*, **20**:611–618.

Wells WF, Stone WR. (1934) On air-borne infection. Study III. Viability of droplet nuclei infection. *Am J Hyg*, **20**:619–627.