Spatial Scan for Disease Mapping on a Mobile Population

Liang Lan
Department of Computer and
Information Sciences, Temple University
lanliang@temple.edu

Vuk Malbasa
Faculty of Technical Science,
University of Novi Sad, Serbia
vmalbasa@gmail.com

Slobodan Vucetic
Department of Computer and
Information Sciences, Temple University
slobodan.vucetic@temple.edu

Abstract

In disease mapping, the spatial scan statistic is used to detect spatial regions where population is exposed to a significantly higher disease risk than expected. In this important application, the current residence is typically used to define the location of individuals from the population. Considering the mobility of humans at various temporal and spatial scales, using only information about the current residence may be an insufficiently informative proxy because it ignores a multitude of exposures that may occur away from home, or which had occurred at previous residences. In this paper, we propose a spatial scan statistic that is appropriate for disease mapping on mobile populations. We formulate a computationally efficient algorithm that uses the proposed statistic to find significant high-risk regions from mobile population's disease status data. The algorithm is applicable on large populations and over dense spatial grids. The experimental results demonstrate that the proposed algorithm is computationally efficient and outperforms the traditional disease clustering approaches at discovering high-risk regions in mobile populations.

Introduction

Disease mapping methods are used to understand the geographic variability in disease risk by studying the association between the occurrence of disease and the locations of individuals in the population. It is an essential tool in modern epidemiology, because location serves as a proxy for lifestyle, social and environmental factors that may be unobserved or unavailable for study. Disease maps have served as a hypotheses generating tool, allowing investigators to draw inferences about disease etiology and make informed decisions about the allocation of public health resources.

There are two major approaches for disease mapping. Both methods require information about location of individuals from the population and their disease status. The first method aims to determine if and how disease risk varies across space. This approach typically relies on computationally expensive Hierarchical Bayesian Modeling (Banerjee, Gelfand, and Carlin 2003; Mollié 1996) to exploit spatial correlation in disease risk. Several Bayesian spatial models have been proposed for disease mapping (Best, Richardson, and Thomson 2005). The method is computationally costly and is effective only when the number of cases (i.e. individuals with a disease) is sufficiently large relative to the spatial resolution. The alternative, called the disease clustering, aims to find spatial regions where there are significantly more cases than what have been expected according to the baseline risk. This widely used approach stems from Kulldorff's spatial scan statistics (Kulldorff 1997). It currently has many variants (Kulldorff et al. 2005; Toshiro and Kunihiko 2005) that can be used for various types of data. The spatial scan has received attention in the machine learning community from the perspective of computational efficiency (Neill and Moore 2004; Neill et al. 2004). Disease clustering is widely applicable because it is robust even when the incidence of disease is relatively low. Disease clustering is the focus of this paper.

The existing disease mapping methods typically use residence of individuals from the population for geo-coding of their location. This can be a serious constraint, considering the mobility of humans at various temporal and spatial scales. At short temporal scales, e.g., at the level of a single day, people typically spend significant time outside of their home doing activities such as work, commuting, entertainment, or travel. At a longer temporal scale, e.g., over years or decades, people typically change residences multiple times. The spatial scale of human mobility can range from a person's movement within a home to intercontinental air travel.

Using only information about the current residence can be misleading because it ignores a multitude of environmental exposures that can occur or have occurred away from the current residence. Let us consider several examples in which the current place of residence is not sufficiently informative: an increased number asthma attacks in people that were at a port while cargo with an allergen was unloaded, a small scale outbreak of the stomach flu among patrons of a downtown restaurant, an increased incidence of lung cancer among people who worked in a particular factory a decade ago. Clearly, information about movement patterns that occurred away from home or at previous residences would be very useful for disease mapping in all of these scenarios.

Until recently, the main obstacle in using mobility data for disease mapping was a lack of technology to collect such data for a significant fraction of a population. However, the

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

almost ubiquitous use of mobile and smart phones, as well as the emergence of geocoded databases about residential histories, makes it possible to obtain detailed and accurate information about mobility of human population at an unprecedented scale and with low-cost. For example, nEmesis-project (Sadilek et al. 2013) developed an intriguing system that analyzes public geocoded tweets from New York City to detect if current reports of foodborne disease symptoms by some users are correlated with their recent visits to particular restaurants. The promising results indicate that it might be possible to utilize public tweets as a useful source of information for disease surveillance. Privacy issues notwithstanding, it is evident that location-based technologies offer a significant opportunity for public health and disease surveillance.

As the mobility data are becoming increasingly available, it is still not clear how to analyze such data to improve quality of disease mapping. In recent years, there have been a few attempts to develop new methods for disease mapping from mobile populations. One is related to the recent interest in the life course approach to health (Pickles, Maughan, and Wadsworth 2007), which emphasizes the significance of timing in associations between physical (e.g., chemical, sun exposure) and social (e.g. poverty, employment) exposures and chronic diseases. Another is development of Q-statistic (Jacquez et al. 2005; Jacquez, Meliker, and Kaufmann 2007), for case-only clustering of movement trajectories which assumes that moving trajectories of cases are grouped over specific spatio-temporal windows, and M-statistic (Manjourides and Pagano 2011), for comparing spatial distribution of cases and controls after weighting historical residences by an assumed incubation time distribution. Both Q- and M-statistics methods are heuristically motivated by spatial scan statistics and use a strong assumption that all cases should have similar movement patterns.

In this paper, we present a novel disease clustering approach which extends Kulldorff's spatial scan statistic to mobility data. Given the information about movement of individuals and their health status, we assume that the probability that an individual becomes sick is a logistic function of a weighted sum of the disease risks at the visited locations. We design a log-likelihood ratio test score and use it to measure if a given sub-region has a significantly higher disease risk than the background risk. We can detect significant sub-regions of any size, located anywhere within the study region. We propose several strategies to reduce the computational cost and make the method applicable to large populations and dense spatial grids. Finally, we show experimental results that demonstrate validity of the proposed approach.

Problem Definition

Let us consider a spatial region inhabited by N individuals and consisting of L locations. We denote the disease status of the i-th individual as $y_i = 1$ if he or she is sick, and $y_i = 0$ otherwise. Let us represent a movement pattern of each individual as the mobility vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iL}]^T$, where x_{il} is the fraction of total time the i-th individual spent at location l ($\sum_{i=1}^{L} x_{ij} = 1$). We denote $\mathbf{r} = [r_1, r_2, \dots, r_L]^T$ as a vector of disease risks, where r_l is a measure of the disease risk of the l-th location. We assume the probability that the i-th individual becomes sick is a logistic function of the weighted average of disease risks at visited locations, $\rho_i = \frac{1}{1+\exp^{-\mathbf{r}^T \mathbf{x}_i}}$. Given the logistic model, the objective of disease mapping is to estimate spatial risks $\mathbf{r}$ from a data set of N individuals, where i-th individual is represented as a pair $(\mathbf{x}_i, y_i)$. This general objective may be too ambitious in the common scenario where the number of cases is relatively small compared to the number of locations. As a consequence, disease mapping often focuses on a simpler problem, called disease clustering, where the objective is to find if there is a sub-region with the statistically significant increased disease risk as compared to the background risk and to find the most significant such sub-region. In this paper, we propose a new method for disease clustering on mobile populations.

Let us denote by r_{in} the risk inside a candidate sub-region R and r_{out} the risk outside the sub-region R. We use $x_{i,in}$ as the fraction of time spent by the i-th individual within sub-region R and $x_{i,out}$ as the fraction of the time spent outside sub-region R. Then, the disease probability for the i-th individual can be expressed as

$$\rho_i = \frac{1}{1 + \exp^{(-(r_{in} x_{i,in} + r_{out} x_{i,out}))}}. \tag{1}$$

For each sub-region R, the objective of disease clustering is to test the null hypothesis $\mathbf{H}_0 : r_{in} = r_{out}$, that disease risks are equal within and outside R. The alternative hypothesis for every sub-region R is $\mathbf{H}_1 : r_{in} > r_{out}$, that the risk within R is higher than the background risk. A challenge is to find an appropriate hypothesis testing strategy that has sufficient power to discover significant sub-regions and do so in a computationally efficient manner. In the following section, we will describe Kulldorff's spatial scan statistic (Kulldorff 1997), which is the most powerful for discovering disease clusters in static population. Then, we will propose how to modify the statistic for finding disease clusters in mobile populations.

Methodology

Original Spatial Scan The Kulldorff's spatial scan (Kulldorff 1997) is appropriate for static population, where it is assumed that individuals spend all their time at their homes. Following the notation introduced in the previous paragraph, the i-th individual is represented by a binary mobility vector $\mathbf{x}_i$ where x_{il} = 1 if location l is the i-th individual's home and x_{il} = 0 otherwise. In Kulldorff's spatial scan, each location is represented with a pair (c_l, p_l), where c_l is the number of cases residing at the l-th location, and p_l is the total number of people residing in the location. For any considered sub-region R, the pairs are summed up to calculate (c_{in}, p_{in}) pair inside the region and (c_{out}, p_{out}) pair outside the region, and a score S_R is calculated as the log of the ratio between two likelihoods,

$$S_R = \log \frac{\max\limits_{\rho_{in},\rho_{out}} P(Data|\rho_{in} > \rho_{out})}{\max\limits_{\rho_{in},\rho_{out}} P(Data|\rho_{in} = \rho_{out})}. \tag{2}$$

The numerator denotes the maximum likelihood of the data under the assumption that the disease probability of an individual in region R (denoted as ρ_{in}) is higher than the disease probability of an individual in the outside region (denoted as ρ_{out}), and the denominator denotes the maximum likelihood of the data under the assumption that the disease risk is identical inside and outside the region. The resulting score of (2) can be expressed as

$$c_{in}\log\frac{c_{in}}{p_{in}} + c_{out}log\frac{c_{out}}{p_{out}} - (c_{in}+c_{out})log\frac{c_{in}+c_{out}}{p_{in}+p_{out}} \quad (3)$$

if $\frac{c_{in}}{p_{out}} > \frac{c_{out}}{p_{out}}$, and 0 otherwise. Kulldorff (1997) proved that this spatial scan score is individually the most powerful for finding a significant region of elevated disease risk.

After the spatial scan scores S_R are calculated for all sub-regions R, the sub-region with the highest score

$$\lambda = \max_R S_R \quad (4)$$

is selected. Since the distribution of the maximal score λ cannot be expressed analytically, to calculate the statistical significance of the sub-region with the maximal score, a costly randomization technique has to be used. There, the disease status labels y_i are shuffled among the N individuals and the maximal score is found on the shuffled data set. This procedure is repeated B times (typically, $B = 100$ or even $B = 1,000$) to produce B maximal scores on B shuffled data sets. If the maximal score on the original data is higher than that on all or a vast majority of shuffled data sets, it can be treated as significant. The ratio between the number of shuffled data sets with the higher score and B can serve as an approximation of the p-value of the null hypothesis that disease risk is constant over the whole region. It should be noted that there are many variants of this procedure with respect to how the score is calculated (Neill 2009). There are also extensions, such as finding the largest spatio-temporal sub-region (Neill et al. 2005) or finding the most significant sub-region for multiple diseases (Kulldorff et al. 2007).

Let us now discuss the computational cost of the described spatial scan approach. Let us assume for simplicity that the whole spatial region can be represented as a squared grid of size $K \times K$ (i.e., $L = K^2$). Since there are $O(K^4)$ rectangular sub-regions within the grid, and $O(1)$ time is enough to calculate the (c, p) pairs for each sub-region, the naive cost of disease clustering using the Kulldorff's method is $O(N) + O(K^4 B)$. The popular SaTScan software for disease clustering discovers only circular sub-regions, which reduces time to $O(N)+O(K^3 B)$. It should be noted that under certain reasonable conditions, including the Kulldorff's spatial scan, and with smart pruning strategies, the time for discovery of rectangular sub-regions could be reduced down to $O(N) + O(K^2 log^2(K)B)$ (Neill and Moore 2004; Agarwal et al. 2006).

Spatial Scan for Mobile Populations We now describe how to develop a spatial scan statistic for disease clustering on a mobile population. Similarly to Kulldorff's spatial scan, we use the likelihood ratio as the test statistic. Let us assume that we are studying sub-region R with disease risks r_{in} within the sub-region and r_{out} outside the sub-region. We can express the likelihood function for a population with N individuals as

$$L(R, r_{in}, r_{out}) = \prod_{i=1}^{N} \rho_i^{y_i}(1-\rho_i)^{(1-y_i)}, \quad (5)$$

where ρ_i is defined in (1). The likelihood ratio is

$$S_R = \frac{\max_{r_{in}>r_{out}} L(R, r_{in}, r_{out})}{\max_{r_{in}=r_{out}} L(R, r_{in}, r_{out})}. \quad (6)$$

When $r_{in} = r_{out} = r$, we can write the likelihood as

$$L(R, r_{in} = r_{out}) = \rho^C(1-\rho)^{N-C}, \quad (7)$$

where $\rho = \frac{1}{1+\exp^{-r}}$, and C is the number of cases in the whole population. The denominator in equation (6) then becomes

$$\max_{r_{in}=r_{out}} L(S, r_{in}, r_{out}) = \frac{C^C(N-C)^{(N-C)}}{N^N} = L_0, \quad (8)$$

because the maximum likelihood is obtained when $\rho = C/N$. Therefore, L_0 is a constant value that depends only on the total number of cases C.

Now, we would like to find the value of the numerator in (6). For a given sub-region R, we need to find the maximum likelihood over all possible $r_{in} > r_{out}$. Instead of maximizing (5), we can maximize the log-likelihood subject to a constraint,

$$\begin{aligned} &\max_{r_{in},r_{out}} \sum_{i=1}^{N}[y_i log(\rho_i) + (1-y_i)log(1-\rho_i)] \\ &s.t.\ r_{in} > r_{out} \end{aligned} \quad (9)$$

After noting that $x_{i,out} = 1 - x_{i,in}$, (9) is equivalent to a constrained logistic regression model with two parameters (i.e., r_{in}, r_{out}) and a single variable (i.e., $x_{i,in}$). The gradient of (9) is

$$\mathbf{g} = \sum_{i=1}^{N}[(y_i - \rho_i)\mathbf{x}_i], \quad (10)$$

and the Hessian of the objective is

$$H = -\sum_{i=1}^{N}[\rho_i(1-\rho_i)\mathbf{x}_i\mathbf{x}_i^T]. \quad (11)$$

The objective function in (9) is concave and a unique global optimal solution can be obtained. The Newton method updates the parameter $\mathbf{r}$ as:

$$\mathbf{r}^{new} = \mathbf{r}^{old} - (H)^{-1}\mathbf{g}. \quad (12)$$

The Hessian matrix is of size 2×2, which allows efficient learning.

Now, let us consider the constraint $r_{in} > r_{out}$. We are only interested in regions R where $r_{in} > r_{out}$. If after solving (9) we get a solution where $r_{in} < r_{out}$, we set the solution to be $r_{in} = r_{out}$, and the corresponding likelihood ratio to 1. Therefore, we can express the log-likelihood ratio for sub-region R as:

$$S_R = \begin{cases} \log \frac{\max_{\mathbf{r}} L_{\mathbf{r}}}{L_0} & if\ r_{in} > r_{out} \\ 0 & if\ r_{in} \le r_{out} \end{cases} \tag{13}$$

Note that if we only use current residence to construct mobility vectors for the individuals, the probability of the i-th individual is $\rho_i = \frac{1}{1+\exp^{-r_{in}}}$ if the i-th individual resides within the sub-region R, and $\rho_i = \frac{1}{1+\exp^{-r_{out}}}$ otherwise. By using the log-likelihood ratio test, S_R from (13) reduces to S_R of the Kulldorff's spatial scan.

Scalability

Trivial Implementation Let us first consider the cost of a trivial implementation of our proposed disease clustering method for mobile populations. For simplicity of the analysis, we assume a $K \times K$ spatial grid with a total of $L = K^2$ locations and the population size of N is given, and we are interested in finding the highest-scoring square subregion R. To obtain the highest score λ, we need to compute S_R for all squares with sizes ranging from $k = 1, \ldots, K$. For any size k, there are $(K - k + 1)^2$ sub-regions. So there are $O(K^3)$ sub-regions to examine. To construct vector $\mathbf{x}_{in} = [x_{1,in}, x_{2,in}, \ldots, x_{N,in}]^T$ needed for logistic regression we need to scan the whole data set, which takes $O(NK^2)$time. Given $\mathbf{x}_{in}$, we need an additional $O(N)$ time to train the model. Therefore, the naive implementation requires $O(K^5N)$ time to compute λ.Since we need to calculate λ values on B shuffled data sets to estimate the statistical significance of the discovered highest-scoring subregion, the total cost becomes $O(K^5NB)$, which is much higher than the cost of the original Kulldorff's spatial scan method for static population. In the following we explain how this trivial cost can be significantly reduced to result in relatively computationally-efficient method that could be applied on large populations with dense spatial grids.

Speedup by Sliding Let us assume that we just examined sub-region $R_{i,j,k}$ of size $k \times k$ starting at position (i, j) on the spatial grid and that we saved its $\mathbf{x}_{in}$ vector.Since the neighboring sub-region $R_{i,j+1,k}$ differs in $2k$ grid cells, only those locations should be scanned to update $\mathbf{x}_{in}$, which takes $O(kN)$ time instead of $O(K^2N)$ in the trivial implementation. Thus, the total time of the method can be reduced to $O(K^4NB)$.

Speedup through Sparsity Mobility vector $\mathbf{x}_i$ of a typical individual is likely to be sparse because a typical individual might only visit a small number of locations during the period of interest. If we denote by s the average number of locations visited by an individual from the population, the average location will be visited by Ns/K^2 individuals. Thus, to update $\mathbf{x}_{in}$ after moving from subregion $R_{i,j,k}$ to $R_{i,j+1,k}$ would take the expected $2kNs/K^2$ time. Thus, calculating $\mathbf{x}_{in}$ for all square sub-regions takes $O(K^2Ns)$ time. By adding the time to train $O(K^3)$ logistic regression models, the total time of the method becomes $O(K^3NB + K^2NsB)$.

Speedup by Discretization The time bottleneck after exploiting the sparsity is in having to train a large number of logistic regression models on $(\mathbf{x}_{in}, \mathbf{y})$ data, which requires $O(N)$ time. Here, we propose a discretization technique to reduce the training set size. Since, the $x_{i,in}$ values are within range [0, 1], we divide the range into M equal bins. The examples with the same discretized value $x_{i,in}$ and label y_i are grouped together. After discretization, the new data set can be represented as $\{x_b, c_b^+, c_b^-\}_{b=1}^M$, where x_b is the corresponding discretized value of the b-th bin, and c_b^+ and c_b^- are the counts of positive and negative examples in discretized bin b. Therefore, (9), (10), (11), (12) can be rewritten as weighted logistic regression,

$$\max_{r_{in}, r_{out}} \sum_{b=1}^{M} [c_b^+ log(\rho_b) + c_b^- log(1 - \rho_b)] \qquad s.t.\ r_{in} > r_{out} \tag{14}$$

$$\mathbf{g} = \sum_{b=1}^{M} [c_b^+ (1 - \rho_b)\mathbf{x}_b + c_b^- (-\rho_b)\mathbf{x}_b], \tag{15}$$

$$H = -\sum_{b=1}^{M} [(c_b^+ + c_b^-)\rho_b(1 - \rho_b)\mathbf{x}_b\mathbf{x}_b^T]. \tag{16}$$

$$\mathbf{r}^{new} = \mathbf{r}^{old} - (H)^{-1}\mathbf{g}. \tag{17}$$

Therefore, the time complexity to solve the weighted logistic regression is $O(M)$. Note the M could be orders of magnitude smaller than N. In our experimental section, we show that setting M to 100 is sufficient to get an accurate solution. The cost to update the discretized version of $\mathbf{x}_{in}$ after moving from sub-region $R_{i,j,k}$ to $R_{i,j+1,k}$ takes the expected $2kNs/K^2$ time. Thus, the total time of the method becomes the appealing $O(K^3MB + K^2NsB)$. If we make the realistic assumption that $N > K$, neglect constants M,B and s, and recall that $L = K^2$, the total cost of the method simplifies to $O(LN)$, which is linear in the population size and number of locations. The similar speedups are possible for rectangular sub-regions, in which case the cost of the proposed method becomes the still acceptable $O(L^{3/2}N)$.

Speedup by Pruning The most common scenario in disease mapping is that cases are only a small fraction of the population. If that is the case, it is possible to further speedup the method by exploiting the fact that most of the locations might not have been visited by cases. Let us consider a case when the score is known for sub-region $R_{i,j,k}$ and that additional locations covered by larger sub-region $R_{i,j,k+1}$have not been visited by cases. Then, it is guaranteed that the score of the larger sub-region cannot be larger than the score of the smaller region. Thus, the score of the larger sub-region does not need to be calculated. With an appropriate bookkeeping, significant savings in computational time could be achieved when number of cases is small.

We note that scalability could be further increased by parallelization, for example by using approach similar to that in our previous work (Djuric, Grbovic, and Vucetic 2013).

Experimental Setting and Results

EpiSims Data In order to evaluate the proposed spatial scan algorithm and to compare usefulness of residential and movement data in detecting significant overdensity clusters, we used EpiSims data set from Network Dynamics and Simulation Science Laboratory (NDSSL 2006). The data set was designed to realistically simulate behavior of the population of Portland, OR, at the level of individual people. This data set contains information about the movement of individuals, the types of their activities, and their social contacts. In particular, this synthetic data set summarizes daily activities of 1,601,329 peoples as they moved within 240,090 locations of the city. For this study, we used only movement trajectories of the individuals.

We processed the original EpiSims data such that the Portland, OR, metropolitan region was partitioned into a regular grid of size 150 × 150, and the original 240,090 locations were assigned to the appropriate grid cells. In the resulting data set, each location was visited by an average of 25 people and each person visited an average of 3 locations. We represented i-th individual by mobility vector $\mathbf{x}_i$, summarizing the fraction of time spent on each grid cell, as explained in the Problem Definition.

In the following experiments, we transformed the 150 × 150 grid into a coarser 50 × 50 grid and pick several square sub-regions as high-risk sub-regions. In each case, we specify r_{in} value within the selected high-risk sub-region and r_{out} for the remaining grid cells. We select r_{in} to be larger than r_{out}. To generate the target y_i for i-th individual, we first compute the probability $\rho_i = \frac{1}{1+\exp^{-\mathbf{r}_i^T \mathbf{x}_i}}$. Then the labels $y_i \in \{0, 1\}$ are generated by throwing a biased coin with this probability. In this way, we generated *the mobility data set* $D_M = (\mathbf{x}_i, y_i), i = 1, \ldots, N$ where $\mathbf{x}_i$ is $L = 150 \times 150$ dimensional vector and $N = 1,601,329$. EpiSims data set also provides information about location of residence for each person. Therefore, we were able to generate another data set, where each person was characterized by a binary mobility vector $\mathbf{x}_i$ where x_{il} = 1 if location l is the i-th person's residence and x_{il} = 0 otherwise. In this way, we generated another data set that we will call *the residential data set* D_R.We note that our proposed spatial scan method is equivalent to the original Kulldorff's spatial scan method on residential data set D_R. Thus, we will be able to directly compare our proposed method with the Kulldorff's method on a number of scenarios.

We need to emphasize that this simulated data set is ideal, because it assumes movement patterns of all individuals are know precisely. In real life, we could expect the data to be incomplete and corrupted, which might require some modifications to the proposed method (Zoeter et al. 2012).

Experiments: Scenario 1 In our first experiment, we used a square with size 3 × 3 centered on "Milwaukie Business Industrial" (denoted as the red solid square in Figure 1) as the high-risk sub-region. This sub-region was chosen because it was the most commonly visited by the simulated population among all squares of that size. We set r_{in} = log(199) and r_{out} = log(999), such that an individual spending all time inside the sub-region would have disease probability $\rho_i = 0.005$, while an individual spending all time outside would have disease probability $\rho_i = 0.001$. In this setting, we randomly sampled $N = 100,000$ people. The selected risks resulted in about 150 generated cases, where about 100 of them did not visit the high-risk sub-region. Then we used our proposed method to detect the most significant sub-region. The detected highest risk sub-regions based on movement D_M data and residential data D_R are shown in Figure 1. The detected sub-region based on the mobility data was of size 6 × 6 and centered across the true high-risk sub-region (shown as the red dotted square in Figure 1). The resulting maximum score λ was 12.17 and it was significantly larger than for any of the $B = 100$ shuffled data sets, indicating that the p-value is below 0.01. The detected sub-region using the residential data was the 11 × 11 black dotted square shown in Figure 1. The resulting maximum score λ was 5.87 and it was higher than the maximum score in only 61 of the $B = 100$ shuffled data sets, indicating the p-value of 0.39.

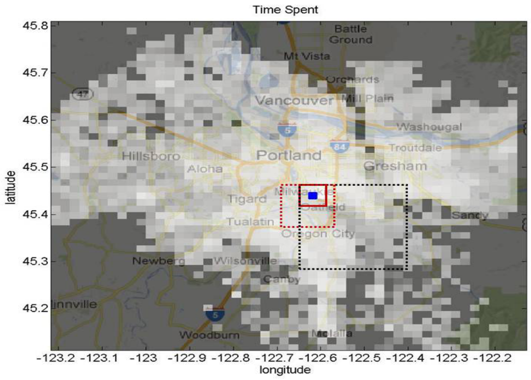

Figure 1: Detected Region for Scenario 1 (*Solid Red Square*: true risk region; *Dotted Red Square*: detected region based on movement data; *Dotted Black Square*: detected region based on static data)

Experiments: Scenario 2 In our second experiment, we selected Portland international airport as the true high-risk sub-region. It was chosen because it is an extreme example of a sub-region visited by many people in which very few people reside. Therefore, only using residential data set is not likely to lead to detection of the high-risk sub-region. In this scenario, we tested our method under several different choices of the disease risk.

Setting 1. In the first case, we set r_{in} = log (199) (i.e. ρ_i = 0.005) and r_{out} = log (999) (i.e. ρ_i = 0.001). We randomly sampled N = 100,000 people from the whole population. We used a square with size 3 × 3 centered on Portland international airport as the true high-risk sub-region(the red dotted square in Figure 2). The detected high-risk sub-regions based on mobility data (dotted red square) and residential data (dotted black square) are shown in Figure 2. The detected sub-region based on movement data was within the true high-risk sub-region, but with p-value of only 0.22. The detected sub-region based on residential data was away from the true high-risk sub-region and its p-value was only 0.49. Thus, neither method returned a statistically significant high-risk sub-region. The reason was that both the disease risk and the size of high-risk sub-region were very small.

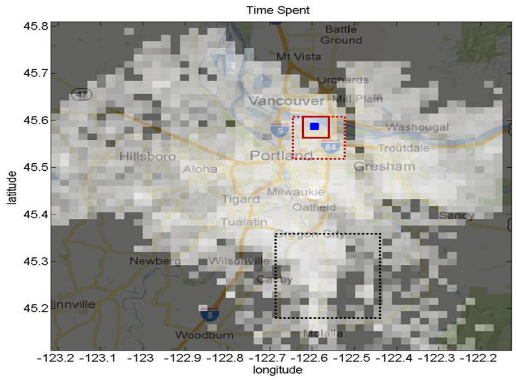

Figure 2: Detected Region for Scenario 2 on Setting 1

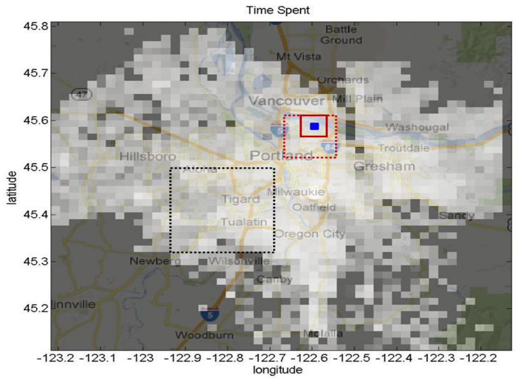

Figure 3: Detected Region for Scenario 2 on Setting 2

Table 1: The Running Time of Proposed Spatial Scan for Different Resolutions K

K	10	30	50	75	150
time (sec)	66	238	715	1438	6153

The actual number of cases in the data set was only 108, where only 9 of them visited the airport sub-region. Such a small number of cases induced by the high-risk sub-region was thus below the sensitivity of the method. However, it should be observed that the highest scoring sub-region contained the actual high-risk sub-region region, so it is possible that this could have been useful information to public health officials.

Setting 2. Here, we slightly increased r_{in} from r_{in} = log(199) (i.e. ρ_i = 0.005) to r_{in} = log(99) (i.e. ρ_i = 0.010). The r_{out} was still fixed at log(999) (i.e. ρ_i = 0.001). In this case, the risk factor difference between r_{in} and r_{out} was somewhat larger and it resulted in 118 cases, and 17 of them visited the airport sub-region. The highest scoring sub-regions based on mobility and residential data are shown in Figure 3. The detected sub-region based on mobility data contained the airport sub-region and had p-value of 0.02. The detected sub-region based on residential data did not contain the airport sub-region and its p-value was not significant at 0.42. We note that, in our experiments in both Scenarios 1 and 2 only the maximum scoring region was significant. The second and lower ranked regions that did not overlap with the highest-scoring region were not significant.

Impact of Spatial Resolution and Discretization In this section, we explore how the computing time depends on the spatial grid resolution (parameter K) and discretization (parameter M). Here we first explored impact of spatial resolution on the computation time. We experimented with the original resolution K = 150 as well as smaller resolutions K = 75, 50, 30, 10. The resulting times are shown in Table 1. As expected, the computing time is near quadratic with respect to the resolution. Second, we explored the impact of data discretization, used data discretization technique to speed up the training time of logistic regression on accuracy and computational time. Let us denote the optimal solution obtained from (9) as r_{opt} and the approximated solution using discretization from (14) as r_{appr}, we used $||r_{opt} - r_{appr}||_2/||r_{opt}||_2$ to denote the solution difference, where $|| \cdot ||_2$ denotes the l_2 norm. In our experimental setting, we increased the number of bins from M = 10 to 10,000, and we fixed $K = 50$. As shown in Figure 4, we got very accurate approximate solution when the number of bins was 100. By increasing the number of bins from 100 to 10,000, the accuracy of log-likelihood estimation improved only slightly (0.03%). The running time increased nearly linearly with M, as shown in Figure 4. We also checked how the discretization impacts the detected regions. Our empirical results show we could get the same detected region and p-value as the original data by setting M to 100. Therefore, by setting the number of bins to 100, we could get a good tradeoff between solution accuracy and running time. Our empirical results also show that the detected region was not changed when M decreased to 10, but its p-value increased above 0.05.

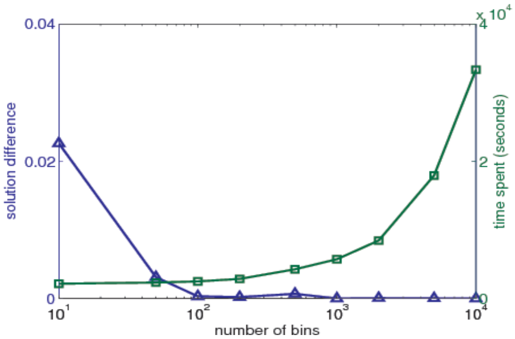

Figure 4: Solution Difference (left y-axis) and Time Spent (right y-axis) based on Different Number of Bins

Conclusion

In this paper, we presented a new test statistic which extends the original spatial scan to movement data. Due to the computational bottleneck of computing the statistic and the significance testing by randomization, an efficient algorithm to compute the spatial scan statistic was proposed. The required computational time is acceptable even for a large population and fine spatial grid resolution. We have performed several experiments to check the difference between using mobility and static data. The experiments clearly show that, if the true risk regions are the locations where few people resided but many people visited, the mobility data are much more useful than residential data. This novel algorithm is very useful for disease monitoring, especially for the environmental diseases (e.g., caner, asthma) where the causative exposures may occurs in the other places which are far away from the individual's current residence. In the future, we would like to further improve the computational efficiency and extend the proposed spatial scan beyond the logistic risk model to cover a larger class of disease models.

Acknowledgements

This work was supported in part by NSF grant IIS-1117433.

References

Agarwal, D.; McGregor, A.; Phillips, J. M.; Venkatasubramanian, S.; and Zhu, Z. 2006. Spatial scan statistics: Approximations and performance study. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 24–33.

Banerjee, S.; Gelfand, A. E.; and Carlin, B. P. 2003. *Hierarchical Modeling and Analysis for Spatial Data*. Crc Press.

Best, N.; Richardson, S.; and Thomson, A. 2005. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14(1):35–59.

Djuric, N.; Grbovic, M.; and Vucetic, S. 2013. Distributed confidence-weighted classification on mapreduce. In *2013 IEEE International Conference on Big Data*, 458–466.

Jacquez, G.; Kaufmann, A.; Meliker, J.; Goovaerts, P.; AvRuskin, G.; and Nriagu, J. 2005. Global, local and focused geographic clustering for case-control data with residential histories. *Environmental Health* 4(1):4.

Jacquez, G.; Meliker, J.; and Kaufmann, A. 2007. In search of induction and latency periods: Space-time interaction accounting for residential mobility, risk factors and covariates. *International Journal of Health Geographics* 6(1):35.

Kulldorff, M.; Heffernan, R.; Hartman, J.; Assuncao, R.; and Mostashari, F. 2005. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2(3):e59.

Kulldorff, M.; Mostashari, F.; Duczmal, L.; Katherine Yih, W.; Kleinman, K.; and Platt, R. 2007. Multivariate scan statistics for disease surveillance. *Statistics in Medicine* 26(8):1824–1833.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26(6):1481–1496.

Manjourides, J., and Pagano, M. 2011. Improving the power of chronic disease surveillance by incorporating residential history. *Statistics in Medicine* 30(18):2222–2233.

Mollié, A. 1996. Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice*. Springer. 359–379.

NDSSL. 2006. Synthetic data products for societal infrastructures and proto-populations: Data set 1.0. *NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, VA, ndssl.vbi.vt.edu/Publications/ndssl-tr-06- 006.pdf.*

Neill, D. B., and Moore, A. W. 2004. Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 256–265.

Neill, D. B.; Moore, A. W.; Pereira, F.; and Mitchell, T. M. 2004. Detecting significant multidimensional spatial clusters. In *Advances in Neural Information Processing Systems*, 969–976.

Neill, D. B.; Moore, A. W.; Sabhnani, M.; and Daniel, K. 2005. Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 218–227.

Neill, D. B. 2009. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 8(1):20.

Pickles, A.; Maughan, B.; and Wadsworth, M. 2007. *Epidemiological Methods in Life Course Research*, volume 1. Oxford University Press.

Sadilek, A.; Brennan, S.; Kautz, H.; and Silenzio, V. 2013. nemesis: Which restaurants should you avoid today? In *First AAAI Conference on Human Computation and Crowdsourcing*.

Toshiro, T., and Kunihiko, T. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4.

Zoeter, O.; Dance, C. R.; Grbovic, M.; Guo, S.; and Bouchard, G. 2012. A general noise resolution model for parking occupancy sensors. In *19th ITS World Congress*.