Spatially Regularized Logistic Regression for Disease Mapping on Large Moving Populations

Vuk Malbasa
Department of Computer and Information Sciences
Temple University
1805 N. Broad St, Philadelphia, PA
vuk.malbasa@temple.edu

Slobodan Vucetic
Department of Computer and Information Sciences
Temple University
1805 N. Broad St, Philadelphia, PA
vucetic@temple.edu

ABSTRACT

Spatial analysis of disease risk, or disease mapping, typically relies on information about the residence and health status of individuals from population under study. However, residence information has its limitations because people are exposed to numerous disease risks as they spend time outside of their residences. Thanks to the wide-spread use of mobile phones and GPS-enabled devices, it is becoming possible to obtain a detailed record about the movement of human populations. Availability of movement information opens up an opportunity to improve the accuracy of disease mapping. Starting with an assumption that an individual's disease risk is a weighted average of risks at the locations which were visited, we show that disease mapping can be accomplished by spatially regularized logistic regression. Due to the inherent sparsity of movement data, the proposed approach can be applied to large populations and over large spatial grids. In our experiments, we were able to map disease for a simulated population with 1.6 million people and a spatial grid with 65 thousand locations in several minutes. The results indicate that movement information can improve the accuracy of disease mapping as compared to residential data only. We also studied a privacy-preserving scenario in which only the aggregate statistics are available about the movement of the overall population, while detailed movement information is available only for individuals with disease. The results indicate that the accuracy of disease mapping remains satisfactory when learning from movement data sanitized in this way.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Apps—Data Mining

General Terms

Algorithms, Performance, Experimentation, Human Factors.

Keywords

Spatial epidemiology, Disease mapping, Movement trajectories, Spatial-temporal data mining, Regularization, Privacy.

1. INTRODUCTION

Disease mapping aims to understand the geographic distribution of a disease by studying the correlation between the occurrence of the disease and the location of individuals from the affected population. It is an indispensable tool in modern epidemiology because location serves as a proxy for detailed lifestyle, environmental, and genetic factors that may be unobserved and unavailable for study. An understanding of the areas with increased disease risk, and the relative ranking of areas according to disease risk, can assist in hypothesis generation about disease etiology and allocation of public health resources.

The origins of disease mapping can be traced back to Finke's 1792 map of indigenous diseases [1] and Snow's cholera maps in the 1850s [2]. In 1854, one of Snow's maps revealed a spatial cluster of cholera cases centered around London's Broad Street pump well, and closing the well resulted in a reduced number of new cases. Further analysis led to the wider acceptance of the germ theory of disease at the end of the 19th century. In Snow's study, geographical information in the form of residential data was used as a surrogate for information about the amount of exposure to contaminated well water. Since then, and especially during the last couple of decades, disease mapping has become a mature technology that has found numerous applications beyond epidemiology, in areas such as psychology, brain imaging, criminology, transportation, forestry, ecology, astronomy, and archeology [3].

There are two major methods for disease mapping. The first, called disease clustering, attempts to find spatial regions whose population has significantly higher disease prevalence than the background population. This widely used approach stems from Kulldorff's work on the spatial scan statistics [4], and it has also received attention in the data mining community [5,6]. The second, more general, method for disease mapping aims to determine the actual spatial map of disease risk. To exploit spatial correlation in disease risk, this approach typically relies on computationally expensive Bayesian hierarchical modeling [7,8]. While the output of disease clustering and disease risk estimation differs, both methods require the same input – information about location and disease status of individuals from the population under study.

The standard approach in disease mapping is to use the place of residence as location information. In this case, individuals are modeled as being affected exclusively by the underlying risk at the place of their current residence. However, due to the ever increasing mobility of modern humans, it is not sufficient to analyze residential data. The dependence of disease risk on movement and exposures outside the primary residence is obvious in infectious diseases; consider a case of food poisoning from a poorly sanitized restaurant in a business district. Similarly, in case of environmental diseases (e.g., cancer, asthma), the disease risk can be a cumulative effect of lifelong exposure to environmental factors (e.g. pesticides, air quality, groundwater pollution) that change over space and time. While residential information is useful as a proxy for an unmeasured aggregate of lifestyle, environmental, and genetic risk factors, it is evident that movement information can provide a more detailed description of these factors and, in many cases, is necessary for successful disease mapping.

Interestingly, there has been limited research on the problem of disease mapping for moving populations. For example, past residences were considered in [9] in order to estimate exposure to chlorination byproducts and their relation to the individual's risk of colon and rectal cancers. In [10], exposure to air pollution was recorded by personal measurement devices and used to provide spatial estimates of pollution. In [11], a spatial scan statistic was applied to a data set of individuals with amyotrophic lateral sclerosis using both place of their current residence and place of birth and it was determined that the place of birth was a better predictor of disease occurrence than the current residence. In [12], a spatio-temporal data mining approach was proposed to analyze residential history information and understand induction and latency periods of individuals with environmental diseases such as cancer.

Historically, a major obstacle towards using movement information in disease mapping was the difficulty in collecting sufficiently detailed movement information about individuals from a population. Recent technological advances, however, make it possible to obtain low-cost and rich information about the movement of people, animals, and goods. This provides an opportunity to improve disease monitoring and to advance knowledge about both infectious and non-infectious diseases. There are two remaining obstacles in achieving this promise: one is the development of novel methods for analysis of movement data, and the other is dealing with privacy issues that could severely restrict access to such data. This paper is a step toward addressing both of these issues.

In this paper we assume that the individual's disease risk is a weighted average of the risks accumulated at visited locations. Given the information about movement of the individuals and their health status, our goal is to uncover the spatial distribution of disease risk within a region of interest. We intend to show that disease mapping can be interpreted as logistic regression from high-dimensional data, where the number of dimensions equals the number of locations. This is an interesting problem because it is characterized by significant sparsity in attributes, but not in the parameters. In contrast, many high-dimensional classification problems, such as text mining and image analysis, are characterized by sparsity of the parameter space. Regardless of this challenge, we demonstrate that disease mapping can be performed efficiently by utilizing the tools of sparse algebra. Spatial regularization, which exploits the spatial correlation in disease risk, is a critical component for the success of disease mapping over large spatial grids. The result is a spatially-regularized logistic regression approach that can be used to perform disease mapping over large or high-resolution spatial regions with thousands of locations from moving populations larger than one million in a matter of minutes.

Privacy is an important aspect of every epidemiological study on human populations. In the case of traditional disease mapping that uses residential information, data are typically aggregated over small areas such that each area has at least some minimum number of residents. Such approaches provide privacy with respect to the popular k-anonymity measure [13]. For movement data, the situation dramatically changes; even for a very coarse spatial aggregation, there might be a significant probability that a large number of individuals have unique moving trajectories [14,15]. Moreover, spatial resolution that is too coarse could negatively influence or completely prohibit accurate disease mapping.

As a step towards addressing the privacy preservation problem in disease mapping on a moving population, we propose a setup that we believe might be acceptable in practice. We assume that exact movement trajectories of healthy individuals are not known, but that the aggregate statistics about total time spent by people at different locations are available. Such statistics are being regularly collected by cell phone companies and shared with their business partners, and they are considered to be sufficiently sanitized to preserve privacy. On the other hand, we assume that precise movement information is available for individuals with disease under a trusted-server setup. In practice, this setup can be achieved by software installed on mobile devices that collects movement information and stores it in an encrypted format on the device. Upon request, the individuals with a disease might be willing to share their movement information with the trusted server maintained by a public health department, under the confidentiality agreement. We will explore experimentally if accurate disease mapping could be achieved under this privacy preserving setup using the modified spatially regularized logistic regression.

2. METHODOLOGY

2.1 Problem Statement

Let us assume we are given a spatial region inhabited by N individuals and consisting of L areas, or locations. For each individual, we have information about the time spent at each of the L locations. In particular, we summarize the movement of the i-th individual with vector $x_i = [x_{i1}, x_{i2} \dots x_{iL}]$, where x_{ij} is the fraction of time the i-th individual spent at the j-th location ($\Sigma_j x_{ij} = 1$). Typically, it is reasonable to assume that an average person visits only a few of the L locations, thus allowing a sparse representation of movement data. Let us 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 also assume that the underlying disease risk can be represented with a vector $\mathbf{r} = [r_1, r_2 \dots r_L]$, where r_j is the measure of risk at the j-th location. It is reasonable to assume that risks are spatially correlated, such that neighboring locations are likely to have a similar risk. For simplicity, in this study we restrict our interest to the static scenario where location risk does not change over time. We also assume that the probability that the i-th individual has a disease is expressed through the weighted average of risks at visited locations,

$$p(y_i \mid \mathbf{x}_i, \mathbf{r}) = \sigma(\mathbf{r}^T \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{r}^T \mathbf{x}_i)}, \tag{1}$$

where σ is the logistic sigmoid function. Given a data set $D = \{(\mathbf{x}_i, y_i),\ i = 1, \ldots, N\}$ of N individuals from the population, the objective is to estimate the underlying disease risk vector $\mathbf{r}$.

2.2 ML and MAP Approaches for Disease Mapping

The Maximum Likelihood (ML) approach for disease mapping consists of maximizing the log-likelihood of training data,

$$\mathbf{r}_{\mathrm{ML}} = \arg\max_{\mathbf{r}} l_D(\mathbf{r}),$$

where

$$\begin{aligned} l_D(\mathbf{r}) &= \ln p(D \mid \mathbf{r}) = \sum_{i=1}^{N} \ln p(y_i \mid \mathbf{x}_i, \mathbf{r}) = \\ &= \sum_{i=1}^{N} \left[y_i \ln \sigma(\mathbf{r}^T \mathbf{x}_i) + (1 - y_i) \ln\left(1 - \sigma(\mathbf{r}^T \mathbf{x}_i)\right) \right] \end{aligned} \qquad (2)$$

This approach leads to logistic regression. The problem with using logistic regression for disease mapping is that the dimensionality of the risk vector $\mathbf{r}$ can be very large. For example, in our experiments we used a regular spatial grid of size 256×256, which resulted in $L = 65{,}536$ locations. In this case, the risk estimates based on ML can be highly unreliable.

To address this issue, we consider a *maximum a posteriori* (MAP) approach that allows us to exploit spatial correlation in disease risk, under the reasonable assumption that disease risk is a relatively smooth spatial process. The MAP approach is based on maximizing the posterior probability $p(\mathbf{r}|D)$, which can be expressed using Bayes Theorem as $p(\mathbf{r}\,|\,D) = p(D|\mathbf{r})\,p(\mathbf{r})\ /\,p(D)$. By observing that $p(D)$ is constant with respect to $\mathbf{r}$, the MAP estimate of risk can be expressed as

$$\mathbf{r}_{\mathrm{MAP}} = \arg\max_{\mathbf{r}} P(\mathbf{r} \mid D) P(\mathbf{r}) = \arg\max_{\mathbf{r}} l_D(\mathbf{r}) + l_P(\mathbf{r}), \qquad (3)$$

where $l_P(\mathbf{r}) = \log P(\mathbf{r})$ is the logarithm of the prior probability of risk. To exploit spatial correlation, we use the Gaussian prior, $\mathbf{r} \sim N(\mathbf{r}|\mu, \mathbf{A}^{-1})$, where μ is a mean vector, typically set to zero, and $\mathbf{A}$ is a precision matrix encoding the neighborhood structure. Typically, $\mathbf{A}$ is sparse and its nondiagonal (k,l)-th element A_{kl} is nonzero only if locations k and l are spatial neighbors. In addition, the diagonal elements of $\mathbf{A}$ are chosen such that each row sums to zero. If this is the case, the prior is called the Gaussian Markov Random Field (GMRF) and $l_P(\mathbf{r})$ can be conveniently represented as

$$l_P(\mathbf{r}) = -\frac{1}{2} \sum_{k \sim l} A_{kl} (r_k - r_l)^2,$$

where $l \sim k$ denotes that l and k are spatial neighbors. Using $l_P(\mathbf{r})$ has the effect of spatial regularization which smoothes the estimated risk, because it encourages neighboring locations to have similar risks. The MAP approach thus leads to spatially-regularized logistic regression.

There are two important questions when implementing this approach: (1) how to choose $\mathbf{A}$ and (2) how to calculate $\mathbf{r}_{\mathrm{MAP}}$ efficiently. Choosing $\mathbf{A}$ consists of defining the spatial neighborhood $N(l)$ of the l-th location and selecting values A_{kl} for $k \in N(l)$. As for the neighborhood, it typically contains only the nearest neighbors (when locations are on a regular grid, these could be the 4 or 8 nearest neighbors), although larger neighborhoods are also possible. A standard way to assign A_{kl} is to set all nonzero non-diagonal elements to the same value λ. In this case, $l_P(\mathbf{r})$ obtains the familiar form

$$l_P(\mathbf{r}) = -\frac{\lambda}{2} \sum_{l \sim k} (r_k - r_l)^2. \qquad (4)$$

As an alternative, if there is strong prior knowledge about spatial covariance $\mathbf{\Sigma}$ (= $\mathbf{A}^{-1}$) of disease risk, one could fit GMRF $\mathbf{A}$ from $\mathbf{\Sigma}$ using recently proposed techniques in [16]. However, we did not pursue this direction further in this study because our preliminary results showed that GMRF fitting is very sensitive and that it does not lead to noticeable improvements in disease mapping accuracy.

Using representation (4), $l_P(\mathbf{r})$ has a single hyperparameter λ that should be determined from training data. In this study, we use a cross-validation (CV) approach to find the best value of hyperparameter λ among a range of possible values. In this case, the training data set is split into several subsets and every subset is used in turn as validation data set to check the accuracy of disease mapping trained on the remaining data for a given choice of λ. We should note that a popular alternative to CV is empirical Bayes, which consists of finding λ as the value that maximizes the likelihood P(D | λ). A very nice overview of this approach that corresponds to regularized logistic regression is given in [17]. However, we did not pursue this approach in our study because the preliminary results suggest that CV is a more robust approach.

Let us now discuss how to obtain $\mathbf{r}_{\mathrm{MAP}}$. The optimization problem (3) is concave and it can be solved using standard tools of convex optimization. In this study we use the Newton-Raphson algorithm due to its ease of implementation and impressive convergence on high dimensional problems. At each iteration of the algorithm, given the current estimate of risk vector $\mathbf{r}^{old}$, it calculates the new estimate $\mathbf{r}^{new}$ as

$$\mathbf{r}^{new} = \mathbf{r}^{old} - \eta \mathbf{H}^{-1} \mathbf{g}$$

where the step size parameter η is often set to 1, $\mathbf{g}$ is the gradient and $\mathbf{H}$ is the Hessian of the MAP function from (3). Specifically, $\mathbf{g} = \mathbf{X}^T(\hat{\mathbf{y}} - \mathbf{y}) - \mathbf{A}\mathbf{r}$ and $\mathbf{H} = \mathbf{X}^T \mathbf{R} \mathbf{X} - \mathbf{A}$, where $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_N]^T$, $\mathbf{y} = [y_1\ y_2\ \ldots\ y_N]^T$, $\hat{\mathbf{y}} = [\sigma(\mathbf{r}^T\mathbf{x}_1) \ldots \sigma(\mathbf{r}^T\mathbf{x}_N)]^T$, and $\mathbf{R}$ is a diagonal matrix with elements $R_{ii} = \hat{y}_i(1 - \hat{y}_i)$. We should observe that although the size L×L of Hessian $\mathbf{H}$ can be extremely large, if $\mathbf{H}$ is sparse $\mathbf{H}^{-1}\mathbf{g} = \mathbf{z}$ can be evaluated efficiently by solving a sparse system of linear equations $\mathbf{H}\mathbf{z} = \mathbf{g}$. Since the precision matrix $\mathbf{A}$ is by design very sparse (it only has few nonzero diagonals encoding the neighborhood) and $\mathbf{R}$ is a diagonal matrix, the sparsity of $\mathbf{H}$ depends only on the sparsity of $\mathbf{X}^T\mathbf{X}$. The (k,l)-th element of $\mathbf{X}^T\mathbf{X}$ is nonzero only if there is a person from the population that visited both the k-th and the l-th locations. When working with high resolution spatial grids, most elements of $\mathbf{X}^T\mathbf{X}$ remain zero, which leads to the sparsity of $\mathbf{H}$.

2.3 Special Case: Disease Mapping on Static Populations

Let us now consider the special case in which the population is static. We show that logistic regression reduces to the standard disease mapping approach. In the case of a static population, each individual from the population is assumed to be affected by the underlying risk of only one location, his or her residence. Let us denote $h(i)$ as the location of the home of the i-th person. In the static scenario, all elements of the movement vector $\mathbf{x}_i$ are zero except $x_{i,h(i)}$ which equals 1. The disease risk of the i-th person equals $p_{h(i)} = \sigma(r_{h(i)})$ and the log-likelihood (2) can be expressed as

$$\begin{aligned} l_D(\mathbf{r}) &= \sum_{i=1}^{N}\left(y_i \log p_{h(i)} + (1-y_i)\log(1-p_{h(i)})\right) = \\ &= \sum_{l=1}^{L} N_{1l} \log p_l + \sum_{l=1}^{L} N_{0l} \log(1-p_l) \end{aligned} \tag{5}$$

where $p_l = \sigma(r_l)$, and N_{1l} and N_{0l} are the numbers of sick and healthy people living at the l-th location. The ML estimation of p_l is obtained from (5) in closed form as the simple ratio

$$p_l = \frac{N_{1l}}{N_{0l} + N_{1l}},$$

which is the ML estimate for the binary distribution. As a consequence, in static populations, counts of the numbers of healthy and sick individuals at a given location are the sufficient statistics for that location. In practice, the ML solution can produce severe overfitting and the MAP approach using the spatial smoothing prior described in Section 2.2 is preferable. Calculating the MAP estimates of risk can be very efficient, as can be seen by observing that $\mathbf{X}^T\mathbf{X}$ becomes a diagonal matrix, which results in an extremely sparse Hessian $\mathbf{H} = \mathbf{X}^T\mathbf{R}\mathbf{X} - \mathbf{A}$.

2.4 Disease Mapping with Privacy

While movement data could certainly increase the quality of disease mapping, privacy concerns are the biggest limiting factor to applying the proposed approach in practice. To address the privacy issue, we will consider a specific privacy scenario that we believe can be practically acceptable. In this scenario, information about movement of the overall population is privacy protected, while the movement information of individuals with disease is available at the trusted server which performs the disease mapping. Our objective is to study how to learn from such data, and to determine what the quality of the resulting disease mapping is.

As a first step, let us rewrite the log-likelihood from (2) as

$$\begin{aligned} l_D(\mathbf{r}) &= l_1(\mathbf{r}) + l_0(\mathbf{r}) = \\ &= \sum_{i=1}^{N_1} \log\sigma(\mathbf{r}^T\mathbf{x}_i) + \sum_{i=1}^{N_0} \log\left(1-\sigma(\mathbf{r}^T\mathbf{x}_i)\right), \end{aligned}$$

where the first term $l_1(\mathbf{r})$ is the log likelihood of N_1 individuals with disease and the second term $l_0(\mathbf{r})$ is log-likelihood of N_0 healthy individuals.

In our privacy-preserving scenario, instead of collecting detailed movement information about healthy individuals, we collect information about the fraction of time t_l the whole population spends at location l, $l = 1, \ldots, L$, and create a vector $\mathbf{t} = [t_1 \ldots t_L]$. Using training data D, t_l can be calculated simply as

$$t_l = \frac{1}{N}\sum_{i=1}^{N} x_{il}\ .$$

Time t_l is the first order statistic that can be collected anonymously, without a need to consult training data. For example, cell phone companies are routinely collecting similar data to optimize their operations or to share data with business partners. By assuming that the majority of the population consists of healthy individuals, statistics $\mathbf{t}$ accurately summarize the spatial distribution of the healthy population.

The learning problem then becomes how to combine statistics $\mathbf{t}$ with detailed movement information about individuals with the disease, represented by a subset D_1 of D, defined as $D_1 = \{(\mathbf{x}_i, 1), i = 1, \ldots, N_1\}$, where N_1 is the size of the subpopulation with the disease. We propose two strategies as described in the following.

Stationary representation. In this approach, we approximate N_0 healthy individuals with a surrogate population of the same size whose individuals spend all time at a single location (e.g. their home), such that there are $N_0 \cdot t_l$ healthy surrogate individuals residing at l-th location. As a result, the aggregated time statistics of the surrogate population equals $\mathbf{t}$. The log-likelihood of the stationary surrogate population is

$$l_0^{Stationary}(\mathbf{r}) = N_0 \sum_{l=1}^{L} t_l \log(1-\sigma(r_l)).$$

An interesting property of the surrogate log-likelihood is that it is a lower bound on $l_0(\mathbf{r})$. This can be seen by using Jensen's inequality and the composition rule as follows. Let us make a reasonable assumption that disease risk is low enough that the probability of disease $\sigma(r_l) < 0.5$ at any location, $l = 1, \ldots, L$. Then, $1-\sigma(\mathbf{r}^T\mathbf{x})$ is a monotonically decreasing concave function. Since $\log(x)$ is monotonically increasing and concave, $\log(1-\sigma(\mathbf{r}^T\mathbf{x}))$ is concave, and so is the log-likelihood l_0. Thus,

$$\begin{aligned} l_0(\mathbf{r}) &> \sum_{i=1}^{N_0}\sum_{l=1}^{L} x_{il} \log(1-\sigma(r_l)) = \\ &= N_0 \sum_{l=1}^{L} t_l \log(1-\sigma(r_l)) = l_0^{Stationary}(\mathbf{r}). \end{aligned}$$

Dispersed representation. In this approach, we approximate each healthy individual with the average individual from a healthy population. In this case, the surrogate population consists of N_0 individuals, each with the moving pattern $\mathbf{x}_i = \mathbf{t}/N_0$. The log-likelihood of the dispersed surrogate population is

$$l_0^{Dispersed}(\mathbf{r}) = N_0 \log(1-\sigma(\frac{\mathbf{r}^T\mathbf{t}}{N_0}))\ .$$

An interesting property of the dispersed log-likelihood is that it is an upper bound on $l_0(\mathbf{r})$. Using the same assumptions and reasoning as for the stationary representation case, we can see that

$$\begin{aligned} l_0^{Dispersed}(\mathbf{r}) &= \log(1-\sigma(\frac{1}{N_0}\sum_{i=1}^{N_0}\mathbf{r}^T\mathbf{x}_i)) > \\ &> \sum_{i=1}^{N_0}\log(1-\sigma(\mathbf{r}^T\mathbf{x}_i)) = l_0(\mathbf{r}). \end{aligned}$$

3. EXPERIMENTAL RESULTS

3.1 Data Set

In order to compare the usefulness of residential and movement data in disease mapping, we used the EpiSims data set from Network Dynamics and Simulation Science Laboratory [18]. This synthetic data set contains information about daily activities of 1.6 million proto-entities as they move around a city with 240 thousand locations. It has been designed to realistically simulate behavior of the population of Portland, OR, at the level of individual people.

The EpiSims data set contains information about movement of individuals, the types of their activities, and social contacts. Previously, we used this information to evaluate a predictor of infection spread [19]. In the present study, we use only movement information for two reasons. First, while today's technology allows seamless collection of movement information, recording activity types and social contacts would require a significant effort and would further complicate the privacy issues. Second, from the perspective of evaluation of the proposed approach for disease mapping, activity types and social contacts are not particularly important.

We processed the original EpiSims data such that the Portland, OR, metropolitan region was partitioned into a regular grid of size 256×256. In the resulting data set, each location was visited by an average of 25 people and each person visited 3 locations on average. In addition to the 256×256 resolution, we also studied coarser grids with 64×64 and 16×16 locations. Each individual was represented by vector **x**, summarizing the fraction of time spent on each of the locations, as defined in Section 2.1. This representation of movement is a sufficient statistic under our assumption that disease risk does not change in time.

To generate targets y, we started by specifying the underlying disease risk map, **r**. Then, for the i-th individual, we calculated the risk $p(y_i|\mathbf{x}_i,\mathbf{r})$ using (1) and generated $y_i \in \{0,1\}$ based on this probability. We call the resulting data set the *movement data set*. In addition to this data set, we generated another one, called the *residential data set*, which contained only residence information (see Section 2.3). This allowed us to examine the benefits of movement data for disease mapping.

In Figure 1, we illustrate the total number of hours (in the $\log_{10}$ scale) the individuals spent at each location according to the movement and residential data sets. As can be seen, the two distributions in the EpiSims data are very similar and there are no obvious regions with significant discrepancies (e.g. regions where few reside, but are visited by many). The only significant discrepancy exists around the Portland airport (see Figure 4).

3.2 Accuracy Measure and Mean Predictor

To calculate the accuracy of disease mapping we used the mean squared error of personal risk prediction,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\sigma(\hat{\mathbf{r}}^T\mathbf{x}_i) - \sigma(\mathbf{r}^T\mathbf{x}_i)\right)^2,$$

where **r** is the true risk, $\hat{\mathbf{r}}$ is the estimated risk, and N is the number of individuals.

In our experiments, in addition to spatially regularized logistic regression using movement and residential data, we also used the baseline mean predictor. This predictor sets the risk at all locations to be constant, with p_l obtained using the ML estimate as

$$p_l = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \forall l,$$

from which the risk r_l can be calculated as $r_l = \log p_l/(1-p_l), \forall l$.

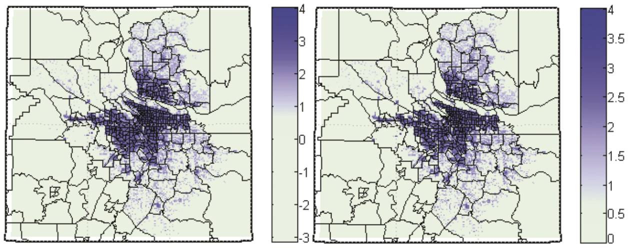

Figure 1. Density of movement (left) and residential (right) data for EpiSims Portland, OR, proto-population

3.3 Experiments: Scenario 1

In Scenario 1 we explore the case where individuals from the studied population reside in a small region, while they visit a significantly larger region. This scenario might be interesting in controlled studies where participants are recruited from a limited geographical region. For this experiment, we sampled $N = 10{,}000$ people living in the Rockcreek and Oak Hills neighborhoods. Figure 2 (left) shows the location of these two neighborhoods and Figure 2 (right) shows the movement density of these people that covers a much larger area.

In our simulations, we placed sources of increased risk at five locations: Reedville, West Heave-Sylvan, Kings Heights, Markham and King City (see Figure 3). We set the background risk to 1% and the peak risk at the 5 sources to 20% in one set of experiments, and to 50% in another. Risk around each of the 5 sources was modeled as Gaussian kernel with width σ. The smallest width ($\sigma = 2$) represented a case of abruptly changing risk where spatial regularization is expected to be less helpful. The largest width ($\sigma = 16$) represented the case with slowly changing risk where even the mean predictor should do well.

In Table 1 we show MSE accuracies for 8 different risk maps, as well as the best value of the regularization parameter λ obtained by cross-validation. We compared accuracies of spatially-regularized logistic regression using movement and residential data, as well as the mean predictor. The results show that using movement data resulted in the highest accuracy in all cases. As expected, the differences were relatively small for estimation of risks generated with the largest and the smallest kernel widths. Interestingly, the accuracies of the mean predictor and spatially-regularized logistic regression based on residential data were almost identical, indicating that, in Scenario 1, the residential data are not very informative. The regularization parameter λ for movement data increased with kernel width, as was expected. It behaved quite erratically with residential data, indicating again the low information content of residential data for disease mapping.

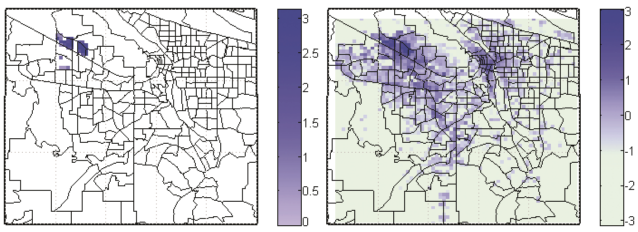

Figure 2. Rockcreek and Oak Hills residential (left) and movement (right) data density

Table 1. Accuracies for Scenario 1

Range	σ	Mean	Residential data		Movement data	
		MSE	MSE	λ	MSE	λ
50% - 1%	2	0.0087	0.0087	500	0.0071	0.02
	4	0.0148	0.0148	200	0.0079	0.05
	8	0.0337	0.0327	0.2	0.0173	0.1
	16	0.0290	0.0270	2	0.0202	0.5
20% - 1%	2	0.0044	0.0044	200	0.0043	0.5
	4	0.0074	0.0074	100	0.0061	0.1
	8	0.0126	0.0124	1	0.0086	0.05
	16	0.0133	0.0132	20	0.0130	5

3.4 Experiments: Scenario 2

In Scenario 2, we focus on the case where increased risk existed within a region in which very few people resided but which was visited by many people. There are many locations such as airports, workplaces, and schools that are visited by a large number of people but are usually not residential zones. In EpiSims data, there is one such region in the vicinity of the Portland International Airport (see Figure 4). For this experiment, we used the whole population (N = 1.6 million). The background risk was set to 1% and the source of increased risk was at the location of airport with maximum of 20% and with small kernel width (σ = 1.6).

As Table 2 shows, the accuracy of disease mapping using movement data is significantly higher than when using residential data or the mean predictor. Similarly to Scenario 1, residential data were not informative, as the mean predictor had the same accuracy. In Figure 4 we illustrate the estimated risk in the vicinity of the airport obtained from movement and residential data. To gain further insight, in Figure 5 we show the scatter plots comparing true risk and estimated risk when movement data (Figure 5.a) and residential data (Figure 5.b) are used. It can be seen that while correlation between actual and estimated risk obtained from movement data is significant, the residential data was not sufficient to uncover the source of increased risk. Therefore, if only residential data are considered, the true source of risk remains hidden and the model reacts by raising the background risk. On the contrary, when using movement data, a more realistic reconstruction of underlying risk can be made.

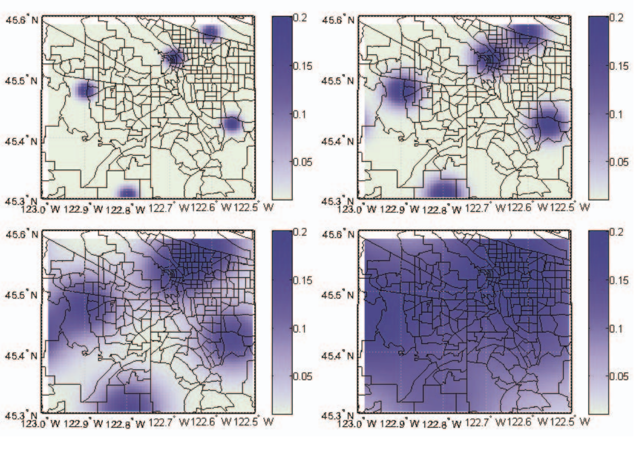

Figure 3. Disease risks in Scenario 1 for different kernel widths

Table 2. Accuracies for Scenario 2

σ	Mean	Residential data		Movement data	
	MSE	MSE	λ	MSE	λ
1.6	0.0041	0.0041	20	0.0019	0.5

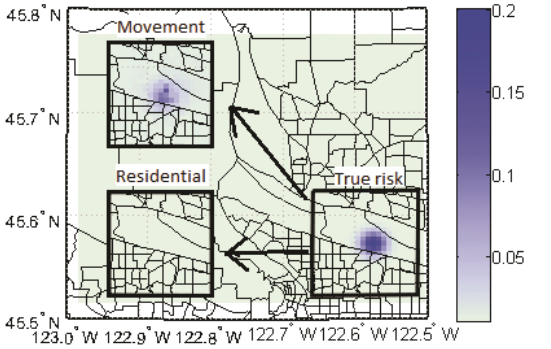

Figure 4. True risk around airport (box on right) and reconstructed risks from movement (top left) and residential (bottom left) data

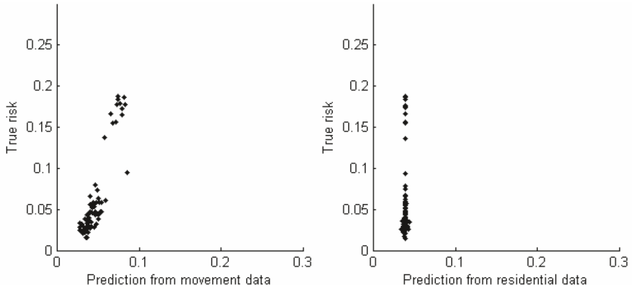

Figure 5. Scatter plots of true risk and risk estimated from movement (top) and residential (bottom) data

3.5 Experiments: Scenario 3

In Scenario 3, we generated spatial risk with 10 sources with kernel width σ = 3 whose peak risks ranged from 10% to 50% (see Figure 6). The background risk was set to 1%. We considered the whole population (N = 1.6 million).

Our first objective was to compare the accuracy of disease mapping when the whole population or only a subset of the whole population (ranging from 0.1% to 50% of the whole population) is observed. From the results in Figure 7, it could be seen that the accuracy when using movement data is consistently higher than when using residential data for all sample sizes. As expected, the MSE decreases with the sample size for both data sets.

Our second objective was to compare accuracy of disease mapping from movement data for different spatial resolutions. In addition to the 256×256 resolution used in Scenarios 1 and 2, we also studied the case when the resolution decreased to 64×64 and 16×16. As could be seen from Figure 8, accuracies using the reduced resolution of 64×64 were comparable to those of the original 256×256 resolution for all sample sizes. On the other hand, the MSE significantly deteriorated when 16×16 was used. In fact, by comparing Figures 7 and 8, it could be seen that for sample sizes of 50% and 100%, the MSE from movement data with 16×16 resolution was higher than that of the residential data with 256×256 resolution.

In Figure 9, we show the time needed to perform disease mapping from movement data for different sample sizes and different data resolutions. As expected, the time increases with sample size and resolution. However, even in the most challenging case with 1.6 million individuals and 256^2 = 65,536 locations, the time needed to estimate disease risk was only 10 minutes. Taken together, the results from Figures 7, 8, and 9 indicate that there is an interesting tradeoff between accuracy and runtime that depends on sample size and spatial resolution.

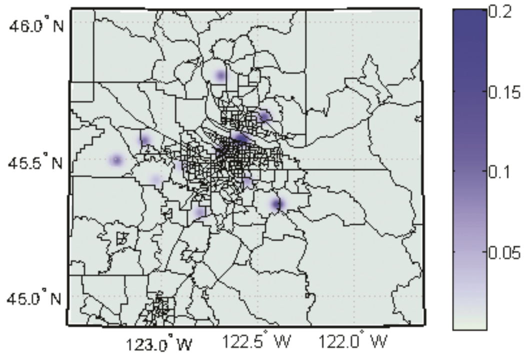

Figure 6. Generated risk for Scenario 3

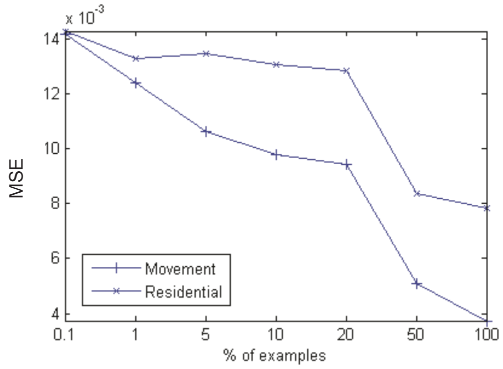

Figure 7. MSE of disease mapping using movement and residential data

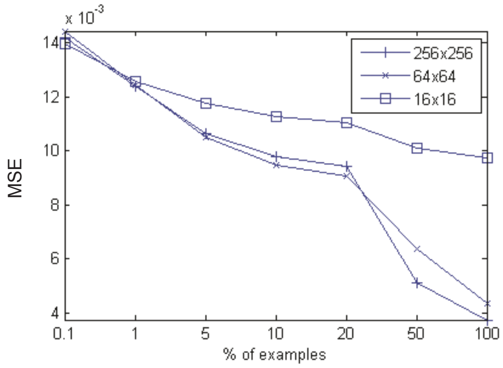

Figure 8. MSE on movement data at varying resolutions

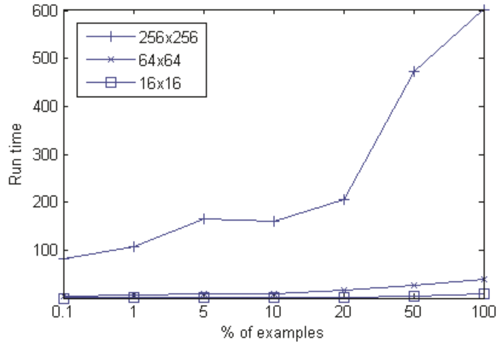

Figure 9: Runtime [in seconds] for different resolutions

3.6 Experiments: Privacy–Preserving Disease Mapping

In the final round of experiments, we evaluated the privacy-preserving strategy proposed in Section 2.4. Using only the aggregated statistics about movement of the whole population and actual movements from individuals with disease, we repeated experiments on data generated in Scenarios 1 and 2. In Tables 3 and 4 we show results only for the stationary representation proposed in 2.4, because in our preliminary results it was observed that using the dispersed representation, the accuracy often dropped below the accuracy obtained using residential data.

For Scenario 1 data, the results shown in Table 3 are directly comparable to the results from Table 1. As can be seen, the stationary representation resulted in accuracy somewhere between the non-sanitized movement data and residential data. We consider this to be a very promising result, as it indicates that the accuracy of traditional disease mapping from residential data can be increased using movement information without seriously infringing on the privacy of individuals. Table 4 shows results for Scenario 2 under the privacy preserving setup. Compared with the results from Table 2, the accuracy (MSE = 0.0030) is just between those obtained using non-sanitized movement data (MSE = 0.0019) and residential data (MSE = 0.0019). It is evident that the difference in accuracy between non-sanitized and sanitized movement data is relatively large. It is an open question if an alternative privacy-preserving approach could be designed that would further improve accuracy of disease mapping and privacy.

Table 3. Accuracies for Scenario 1 – privacy-preserving data

Range	σ	Stationary data	
		MSE	λ
50% - 1%	2	0.0083	0.05
	4	0.0126	0.02
	8	0.0265	0.05
	16	0.0241	0.5
20% - 1%	2	0.0043	0.5
	4	0.0069	0.2
	8	0.0109	0.1
	16	0.0130	5

Table 4. Accuracy for Scenario 2 – privacy-preserving data

σ	Stationary data	
	MSE	λ
1.6	0.0030	0.5

Finally, in Figure 10 we illustrate how accuracy changes as a function of hyperparameter λ in Scenario 2. It can be observed that the accuracy is a well-behaved smooth function of λ that indicates that the search for the best λ should be easy. It is also interesting to observe that the best λ for sanitized and non-sanitized movement data is very similar, while large λ appears to be a good choice for residential data.

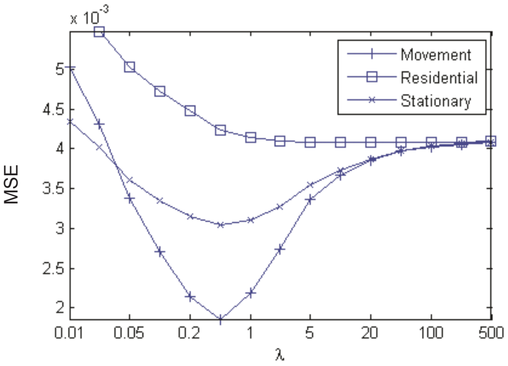

Figure 10. MSE as a function of λ for Scenario 2 data

4. CONCLUDING REMARKS

In this paper we have illustrated that movement data can be very useful for disease mapping as an alternative to the traditional residence data. Under the assumption that the personal risk is a weighted average of risks at visited locations, we showed that disease mapping from movement data can be solved by spatially regularized logistic regression. We also showed that traditional disease mapping is a special case of our approach, where each individual is assumed to be static and spending all of his or her time at the place of residence. By acknowledging privacy issues related to collecting movement information from the whole population, we proposed a privacy-preserving alternative that only requires knowledge of the movement of individuals with disease. The results showed that this setup still allows more accurate disease mapping than when using residential data.

Our study should be considered a step toward exploiting movement data for improving disease mapping. There are numerous open questions to be addressed in future research, of which we mention some. First, our assumption that the personal risk is a weighted average of risks from visited locations from eq. (1) is just one of the possible scenarios. For example, in some cases it might be more reasonable to assume that the personal risk is the maximum risk from the visited locations. In other scenarios, it could be more appropriate to consider the total time spent at different locations instead of the fractional time. Second, our assumption that spatial risk is static may be unrealistic in many disease mapping scenarios. If the risk is allowed to change in time, disease mapping would require estimating the evolving risk map. Third, in some cases, demographic information might be available in addition to the movement data. Integrating the two sources of information for disease mapping is an interesting open problem. Fourth, there is an open question of what is the best choice of the spatial prior, with respect to accuracy of disease mapping, prevention of overfitting, and computational costs. Fifth, our results indicate that spatial variation of disease risk is closely tied with the choice of the spatial prior and spatial grid resolution. It would be interesting to consider how to exploit this relationship to improve both accuracy and computational costs. Sixth, while we believe that the proposed privacy preserving scheme is reasonable, the privacy issue is evidently the main limiting factor in using movement data for disease mapping. There is more space to design improved solutions, both with respect to accuracy and maintaining privacy.

5. ACKNOWLEDGEMENTS

This work was supported in part by the U.S. National Science Foundation Grant IIS-0546155. We thank Madhav Marathe from Virginia Bioinformatics Institute for providing access to EpiSims data to data mining community. We thank Vladan Radosavljevic from Temple University for his valuable advice. We also thank Gregory Johnson from Temple University for his help during the early stages of this study.

6. REFERENCES

[1] F. A. Barrett. Finke's 1792 map of human diseases: the first world disease map? *Social Science & Medicine*, 50:915-921, 2000.

[2] J. Snow. *On the mode of communication of cholera*. Churchill, London, UK, 1855.

[3] M. Kulldorff. *SaTScan^{TM} User Guide*. Available at: http://www.satscan.org [accessed 25 May 2011].

[4] M. Kulldorff. A spatial scan statistic. *Communications in Statistics – Theory and Methods*, 26:1481-1496, 1997.

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

[6] D. B. Neill, A. W. Moore. Anomalous spatial cluster detection. In *Proceedings of KDD 2005 workshop on data mining methods for anomaly detection*, 41-44, 2005.

[7] A. Mollié. Bayesian mapping of disease. In *Markov chain Monte Carlo in practice,* W. R. Gilks, S. Richardson, D. J. Spiegelhalter eds. 359-379, Chapman and Hall, London, UK, 1996.

[8] S. Banerjee, B. P. Carlin, A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, London, UK, 2003.

[9] M. E. Hildesheim, K. P. Cantor, C. F. Lynch, M. Dosemeci, J. Lubin, M. Alavanja, G. Craun. Drinking water source and chlorination byproducts II. Risk of colon and rectal cancers. *Epidemiology*, 9:29-35, 1998.

[10] E. Nethery, S. E. Leckie, K. Teschke, M. Brauer. From measures to models: an evaluation of air pollution exposure assessment for epidemiological studies of pregnant women. *Occupational and Environmental Medicine*, 65:579-586, 2008.

[11] C. E. Sabel, P. J. Boyle, M. Loyonen, A. C. Gatrell, M. Jokelainen, R. Flowerdew, P. Maasilta. Spatial clustering of

amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology*, 157:898-905, 2003.

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

[13] L. Sweeney. k-Anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557-570, 2002.

[14] M. Terrovitis, N. Mamoulis. Privacy preservation in the Publication of Trajectories. In *MDM'08: Proceedings of The Ninth International Conference on Mobile Data Management*, 65-72, 2008.

[15] M. E. Nergiz, M. Atzori, Y. Saygin. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the 2nd SIGSPATIAL ACM GIS International Workshop on Security and Privacy in GIS and LBS,* 52-61, 2008.

[16] H. Rue, H. Tjelmeland. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29:31-49, 2002.

[17] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, R. B. Rao. Bayesian multiple instance learning: Automatic feature selection and inductive transfer. In *Proceedings of the 25th International Conference on Machine learning*, 808-815, 2008.

[18] 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

[19] S. Vucetic, H. Sun. Aggregation of location attributes for prediction of infection risk. In *Proceedings of SIAM DM 2006 Workshop on Spatial Data Mining: Consolidation and Renewed Bearing*, 2006.