Incorporating data on residential history for disease mapping

Caroline Jeffery [*] Justin Manjourides [†] Al Ozonoff [‡] Marcello Pagano[§]

Abstract

When studying the relationship between an individual's location and the acquisition of disease, the location to use is not always clear. When location at exposure is different from that at diagnosis, the latter may not represent the relevant information. While time and location of exposure are often unknown, residential history of cases can substantially inform a spatial analysis. In spatial surveillance, spatial data on cases are often used to detect and locate subareas of the study region with higher or lower risk of disease. Current literature has adapted detection methods to incorporate residential history of cases where available. We extend a disease mapping method to incorporate such data. Through simulations we show that our method is more accurate at identifying a localized increased risk of disease when compared to mapping when only location at diagnosis is considered.

Key Words: spatial surveillance, disease mapping, distance-based method, residential history, cluster detection, incubation distribution

1. Introduction

When public health officials are to make a decision, they rely on prior information assessing the state of health of a particular population. Such information can consist of a health event (e.g. occurrence of female breast cancer) within a study region (e.g. in the state of Massachusetts), a study period (e.g. during 1980-1990), as in [1], and possibly in some subgroup (e.g. for women with age greater than 55). Establishing this knowledge properly requires collecting accurate information and developing precise analytical tools. In this work we consider methods that quantify a variable across a region. Examples of such variables in the context of public health are measures of disease occurrence, environmental exposure, access to care or socio-economic characteristics. Knowing how the quantity varies throughout the region allows for targeted decisions and interventions. When focusing on disease surveillance or syndromic surveillance, the development of these methods promotes using spatial data as an important source of information [2, 3].

Suppose we have a collection of cases diagnosed for a particular disease during a fixed time period, with a location available for each of them. Typically, this collected information refers to where the cases were diagnosed or where they lived at time of diagnosis, and usually falls within a fixed study region. This collection of points represents a sample from a spatial distribution from which to draw inference. In public health surveillance, inference questions are usually framed to determine whether this spatial distribution is unusual or not. Statistical methods can be developed to answer the following questions: At a given time point, is the spatial distribution of cases as expected? If not, are there specific areas in the region with an excess or lack of cases? While the first can be answered globally or locally, the second specifically calls for local approaches. Quantifying locally how unusual the spatial

[*] International Health Group, Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK

[†] Department of Biostatistics, Harvard School of Public Health, Boston MA 02115, USA

[‡] Biostatistics Core, Clinical Research Program, Children's Hospital, Boston 02115 MA, USA

[§] Department of Biostatistics, Harvard School of Public Health, Boston MA 02115, USA

distribution of cases is compared to a known reference distribution provides an example of a variable measured across the study region, specifically a local measure of disease risk. We will refer to such an approach as *disease risk mapping.*

While a single location is often the only spatial information available, it describes individuals as static rather than mobile. However in the same way that a person's age or blood pressure changes overtime, individuals change locations throughout the course of a day or their lifetime [4]. Hence the location where individuals contract diseases can be very different from their location at diagnosis. For some diseases with long latency periods, the time between exposure and diagnosis can span many years [5]. Reporting cases' living address at diagnosis may not provide useful spatial information for studying any geographical exposure, unless it is restricted to those who have resided at their address at diagnosis for many years [1]. When the exposure is related to a particular area of the study region, incorporating the residential history in a quantitative analysis can have several consequences. First, if cases have moved away from the area where they were exposed, the contrast between the exposure area and the remaining part of the study region is attenuated. Second individuals selected to represent the reference population (e.g. population at risk) whose reported address falls in the exposure area further diminish this contrast. Both situations will reduce any spatial method's ability to pick up the existence of an area at risk. Hence it is important to consider analyzing a history of residential addresses rather than just the location at diagnosis. Furthermore, many exposures are positively related to risk of disease, that is a longer exposure increases the risk. Alternatively, some risks sources are only in existence for specific periods of time [6]. Having the duration of stay along with the residential history can therefore lead to even more informative analysis.

A few authors have considered extending spatial methods to incorporate residential history. Jacquez [7] provides a detailed description on how to adapt global, local and focused geographic tests for clustering, which are variations of Cuzick and Edwards's nearest neighbor method [8]. Recently Manjourides and Pagano [9, 10] have developed an extension for the M-statistic, an approach originally designed as a global test of clustering from the perspective of the distribution of distances between points [11]. Global clustering methods have an advantage in terms of power of detection, since they consists of a single test, however they do not locate where any clustering might occur. Focused methods on the other hand require choosing the focus of possible clustering. Finally, the local approach extended in [7] investigates clustering at any case location in the region, but the methods does not account for multiple testing. As an alternative methodology to studying the spatial information on cases, disease mapping methods focus on estimating a change in the spatial distribution of cases from a reference distribution across the study region. Currently, however, they only apply to a single location per case.

In this work we propose to adapt a disease mapping method [12, 13] to incorporate residential history. In the next section, we present a detailed description of this new approach based on the ideas of Manjourides and Pagano [9, 10]. The method is evaluated by simulations in the unit square in the third section. We finish with some general comments about the advantages and limitations of our proposed approach.

2. Methods

2.1 Notation and Definition

Suppose cases are located within a fixed study region R, subset of $\mathbb{R}^2$, and let $\|.\|$ be the Euclidean distance measure defined on $\mathbb{R}^2$. We assume the region R is the support for both the locations and the domain of the mapping function. In practice the latter is represented by a finite set of *grid points* $\{y_1, \ldots, y_r\}$ in R, which are chosen by superimposing a lattice over the study region. Hence we define the mapping function on $R' = \{y_1, \ldots, y_r\} \subset R$.

Definition 2.1. Let $R' = \{y_1, \ldots, y_r\} \subset R$ be a finite collection of points in the Euclidean space. We call any real-valued function $M(.) : R' \to \mathbb{R}$ a *(disease) mapping function*, and we call the range of M, i.e. the set of function values $M(y_1), \ldots, M(y_r)$, the set of *(disease) scores*.

In the context of mapping a risk of disease, we can define the sample space and the mapping function more precisely. Assume that the spatial data on each case consists of a single location in R, then this location is a random point in R. In practice, since the population from which cases arise is finite, the sample space Ω should be a finite collection of points in R. However to allow for a flexible framework, such as defining probability density functions (PDF) and incorporating the mobility of individual over time, we let $\Omega = R$.

Consider now that X is the bivariate random vector representing the location of a case in $\Omega = R$ and let $F : \mathbb{R}^2 \to \mathbb{R}$ be the corresponding bivariate cumulative distribution function (CDF). We frame a disease risk mapping function as a comparison between F and a reference function $F_0 : \mathbb{R}^2 \to \mathbb{R}$ throughout the region R, where the two functions represent respectively the 'observed' and 'expected' spatial distribution of cases.

2.2 Distance-based mapping (DBM) for a single location per case

Standard risk mapping approaches are based on kernel density estimates, which suffer from the curse-of-dimensionality if the methods are applied to dimensions higher than $\mathbb{R}^2$ [14, 15]. The general motivation behind the mapping approach we have proposed previously [12, 13] is to avert the curse-of-dimensionality by making a comparison of the observed and expected spatial distributions in one dimension. Any potential loss of information from the projections is recovered by considering multiple projections and combining the different comparisons. Similarly to tomographic imaging, the two-dimensional space is studied as a fixed number of one-dimensional slices.

We [12, 13] place a fixed number of 'circle points' along a circle circumscribed to R. Each of the N circle points c_i governs a projection, where the one-dimensional counterpart of X is defined as the distance $\|c_i - X\|$, with associated CDF $F_i(t) = \int_{\|c_i - x\| \leq t} dF(x)$. A one-dimensional CDF, F_{0i}, is also defined based on F_0. Then a one-dimensional comparison of F_i and F_{0i} is constructed as a function $\gamma_i : \mathbb{R} \to \mathbb{R}$, where for any real number t and a neighborhood $\mathcal{N}_i(t) = \Big(t - \frac{h(t,i)}{2}, t + \frac{h(t,i)}{2}\Big)$:

$$\gamma_i(t) = \psi\Big(\int_{\mathcal{N}_i(t)} dF_i, \int_{\mathcal{N}_i(t)} dF_{0i}\Big).$$

The comparison function ψ is usually chosen as the difference between the two integrals, but other functions such as a weighted difference or a ratio are possible. The width $h(t,i)$ of the neighborhood is selected so that the integral remains fixed under the null: for a proportion p_0, we [13] define $h(t,i)$ as the unique solution to $\int_{\mathcal{N}_i(t)} dF_{0i} = p_0$. The parameter p_0 acts as a smoothing value and remains fixed for any i and t. This common feature to all these projections guarantees that they all play comparable roles. The final disease mapping function $\Gamma : R' \to \mathbb{R}$ is defined for any point y in the finite region as the average of the one-dimensional comparisons around $\|c_i - y\|$:

$$\begin{aligned}\Gamma(y) &= \frac{1}{N}\sum_{i=1}^{N}\psi\Big(\int_{\mathcal{N}_i(\|c_i-y\|)} dF_i, p_0\Big)\\ &= \Big(\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{N}_i(\|c_i-y\|)} dF_i\Big) - p_0 \qquad \text{if } \psi \text{ is the difference function,} \qquad (1)\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}_F\Big(I\big(\|c_i - X\| \in \mathcal{N}_i(\|c_i-y\|)\big)\Big) - p_0, \qquad (2)\end{aligned}$$

where $\mathcal{E}_F$ denotes the expected value relative to F and $I(\mathcal{S}) = 1$ if $\mathcal{S}$ is true and 0 otherwise. Since p_0 remains fixed, expressions (1) and (2) show that the mapping function Γ is comparing a transformed version of F to p_0. When the observed and expected are the same, Γ equals a scalar throughout the region and we say the map is flat.

To define an estimator for Γ, assume F_0 remains known and suppose the locations of n cases are represented by i.i.d. random variables $X_1, \dots, X_n$, distributed according to F. For each projection we define the one-dimensional empirical cumulative distribution function (ECDF) as $\hat{F}_i(t) = \frac{1}{n}\sum_{j=1}^{n} I(\|c_i - X_j\| \le t)$. Then for fixed $y \in R'$ a consistent estimate of $\Gamma(y)$ in (1) is

$$\widehat{\Gamma}(y) = \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{N}_i(\|c_i-y\|)} d\hat{F}_i - p_0.$$

2.3 Disease mapping for residential history

2.3.1 Example of data

Suppose we now allow individuals to move within the study region R over time. We can describe the location of an individual in time by a pair (A,T), where A is a point in the study region and T is the first time at which the individual resides at that location. The distribution of this three dimensional random vector is represented by a CDF $F_{A,T} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$. Suppose we observe cases of a particular disease diagnosed during a fixed time period $[\tau_{\mathsf{s}}, \tau_{\mathsf{e}}]$, where τ_{s} and τ_{e} respectively mark the start and end times of the collection period, and assume we know the residential history of cases during a period of length τ before they were diagnosed. That is, if T_{d} denotes the time at diagnosis of a selected case, we assume that the complete address history of that case during the time period $[T_{\mathsf{d}} - \tau, T_{\mathsf{d}}]$ is available. Define also the random variable E on $\mathbb{N}^*$ such that $E - 1$ denotes the number of times a case has moved locations during $(T_{\mathsf{d}} - \tau, T_{\mathsf{d}})$. We can represent the information available for each case by the following random vector:

$$\tilde{X} = \{T_\mathsf{d}, E, (A_1, T_1), (A_2, T_2), \dots, (A_E, T_E) : T_1 \leq T_\mathsf{d} - \tau < T_2 < \cdots < T_E < T_\mathsf{d}\}. \quad (3)$$

The CDF $F_{\tilde{X}}$ of $\tilde{X}$ is a real valued function defined on $\mathbb{R} \times \mathbb{R} \times \mathcal{P}(\mathbb{R}^2 \times \mathbb{R})$. Figure 1 gives an example of such spatio-temporal information: a case is diagnosed in 2005, and its residential history is available for a period of $\tau = 20$ years prior; this individual has lived in $E = 4$ different locations, for which three of the starting times fall during the study period $[1985, 2005]$, but the starting time of the earliest location is in 1980, which lies outside the study period.

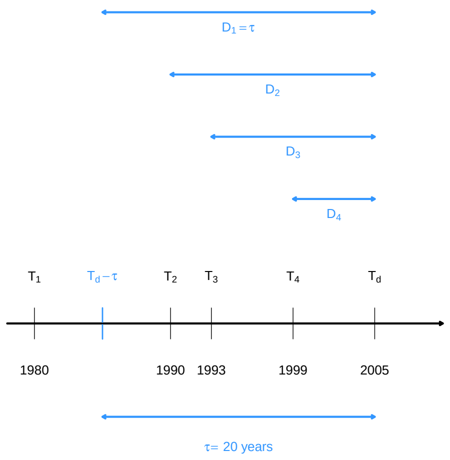

Figure 1: Example of starting times for a case with a history of four residences during the 20 years before diagnosis (Section 2.3.1) and corresponding durations (Section 2.3.2).

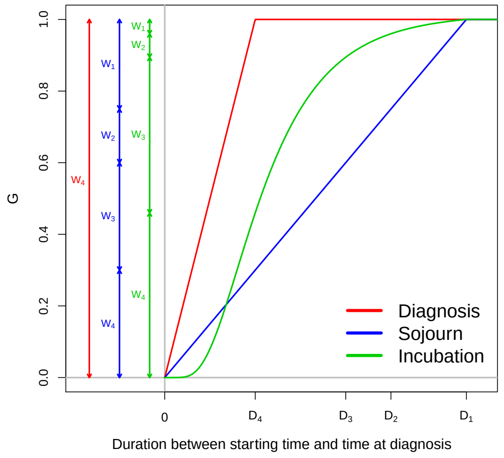

Figure 2: Three examples for the CDF G: All weight given to location at diagnosis (Red), weights based on time spent at location (Blue), weights based on incubation distribution (Green).

2.3.2 *Creating Weights*

We use the different starting times as a basis for defining weights. For each location, define a duration:

$$D_k = min(\tau, T_\mathsf{d} - T_k) \text{ for } k = 1, \dots, E.$$

Except for possibly the earliest location A_1, the random variable D_k measures the length of time between the start of residence at location A_k and the time of diagnosis. We also define $D_{E+1} = 0$. Given the ordering of the starting times, we have $D_{E+1} = 0 < D_E < D_{E-1} < \cdots < D_1 = \tau$. Figure 1 illustrates how to calculate the duration for our example with four available locations during the time period [1985,2005]. Given any CDF $G : \mathbb{R} \to [0, 1]$, we now define the following weights:

$$W_k = G(D_k) - G(D_{k+1}) \text{ for } k = 1, \dots, E.$$

Since G is a CDF, the sum of the weights is $\sum_{k=1}^{E} W_k = G(D_1) = G(\tau) = 1$.

Examples We can define the CDF as $G(v) = \int_0^v g(u)du$ and choose one the following functions $g : \mathbb{R}^+ \to \mathbb{R}$:

- Diagnosis: $g(u) = \frac{1}{D_E} I\big(u \in [0, D_E]\big)$. This gives all weight to the location at diagnosis.
- Sojourn: $g(u) = \frac{1}{\tau} I\big(u \in [0, \tau]\big)$. This defines the weights as the proportion of time spent at each location.
- Incubation: $g(u) = g_{\mathcal{I}}(u) / \int_0^\tau g_{\mathcal{I}}(u) du$ where $g_{\mathcal{I}}$ is the incubation PDF of the disease. This defines the weights as the probability that the lag between exposure and diagnosis falls between two successive durations D_{k+1} and D_k.

Figure 2 illustrates how these three examples impact the resulting weights. The first example gives a positive weight only for the most recent location and disregards all others. The function depends on the duration D_E, hence it will vary from one case to another. The second and third examples on the other hand give a non-zero weight to all locations.

2.3.3 How is DBM adapted to residential history?

As defined in Section 2.1, a disease mapping function assigns a value to each point $\{y_1, \ldots, y_r\}$ in the study region. It aims to assess whether the risk of being diagnosed during the period $[\tau_{\mathsf{s}}, \tau_{\mathsf{e}}]$ varies across the region based on individual's spatial locations over the last τ time units (e.g. years).

To characterize the mapping for multiple locations, we make two assumptions about the components of $\tilde{X}$. First, the number E of pairs (A, T) recorded for a case only depends on the time of diagnosis and τ. Furthermore, given T_{d}, τ and E, these pairs of location and starting time are independent of each other, except for the time ordering in (3). In particular, the durations between starting times are independent. Hence, we can write the probability density function (PDF) of $\tilde{X}$ as

$$\begin{aligned} &f_{\tilde{X}}(t_{\mathsf{d}}, e, a_1, t_1, \ldots, a_e, t_e | \tau_s, \tau_e, \tau) = f_{\tilde{X}}(t_{\mathsf{d}}, e, a_1, t_1, \ldots, a_e, t_e) \\ &= \alpha^{-1} f_{T_{\mathsf{d}}}(t_{\mathsf{d}}) f_{E|T_{\mathsf{d}}}(e) \left\{ \prod_{k=1}^{e} f_{A,T}(a_k, t_k) \right\} I[t_1 \leq t_{\mathsf{d}} - \tau < t_2 < \ldots < t_e \leq t_{\mathsf{d}}], \end{aligned} \tag{4}$$

where $f_{T_{\mathsf{d}}} : [\tau_s, \tau_e] \to \mathbb{R}$ is the PDF of T_{d} and $f_{E|T_{\mathsf{d}}} : \mathbb{N}^* \to \mathbb{R}$ is the probability mass function (PMF) of $E|T_{\mathsf{d}}$. The scalar α is a normalizing constant that guarantees $f_{\tilde{X}}$ integrates to one.

Based on Section 2.3.2 we now have a weight W_k attached to each A_k and $\tilde{X}$ has been transformed as $\tilde{Y} = \{E, (A_1, W_1), \ldots, (A_E, W_E)\}$ with CDF $F_{\tilde{Y}} : \mathbb{R} \times \mathcal{P}(\mathbb{R}^2 \times \mathbb{R}) \to \mathbb{R}$. The starting times are now encoded in the weights.

We propose to define our mapping for multiple locations as a comparison of $F_{\tilde{Y}}$ to a pre-specified reference population $F_{0\tilde{Y}} : \mathbb{R} \times \mathcal{P}(\mathbb{R}^2 \times \mathbb{R}) \to \mathbb{R}$ whose associated PDF $f_{0\tilde{X}}$ can be decomposed in a similar fashion to (4). The one-dimensional comparisons are now constructed as weighted sums. Based on (2), for $y \in R'$, we define:

$$\Gamma(y) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_{F_{\tilde{Y}}} \Big(\sum_{k=1}^{E} W_k I\big(\|c_i - A_k\| \in \mathcal{N}_i(\|c_i - y\|)\big) \Big) - p_0. \tag{5}$$

The width of the intervals $\mathcal{N}_i$ are determined so that for any i,

$$\mathcal{E}_{F_{0\tilde{Y}}} \Big(\sum_{k=1}^{E} W_k I\big(\|c_i - A_k\| \in \mathcal{N}_i(\|c_i - y\|)\big) \Big) = p_0. \tag{6}$$

2.4 Simulations

To evaluate the performance of our adapted mapping function in Section 2.3, we consider simulated data for cases with residential history in the unit square [9, 10]. Individuals are allowed to move within the square during a fixed time study period $[0, \tau]$. For simplification, we assume all individuals are recruited at time $T_{\mathsf{d}} = \tau$. A certain proportion of cases have their disease status linked to a small portion of the square via one of their locations, hence dichotomizing the study region in a high risk area and a low risk area. We give further details in the sections below.

2.4.1 Reference population $F_{0\tilde{X}}$

The residential history distribution of a reference population depicts the underlying distribution from which we want to distinguish the distribution of the residential history of cases. The fixed period for which it is available is also $[0, \tau]$. For this underlying distribution, conditional on the total numbers of locations per individual during the study period, we assume that location and starting time at that location to be independent. Locations are distributed uniformly in the unit square $R = [0, 1] \times [0, 1]$. Starting times are generated such that the time length between two successive starting times follows an exponential distribution $Exp(\lambda)$. More precisely we define the starting times the following way:

$$\begin{aligned} T_1 &= 0 \\ T_2 &= T_1 + min(Exp(\lambda), \tau - T_1) \\ &\vdots \\ T_E &= T_{E-1} + min(Exp(\lambda), \tau - T_{E-1}) \\ T_{E+1} &= T_E + min(Exp(\lambda), \tau - T_E) = \tau, \end{aligned}$$

so that T_{E+1} marks the end of the available residential history. The parameter $1/\lambda$ is the average length of time between successive address changes.

2.4.2 Case population $F_{\tilde{X}}$

The residential history of cases is distributed in a similar fashion to the reference population, however we create an association between one of the locations and the duration spent at that location. First we select a proportion q_0 of cases for which this association will occur. We also select a circular region $\mathcal{C}$ in the unit square, where the center is drawn uniformly in R and the radius is fixed to 0.1 units. This subregion will be the portion of the study region where some of the cases will have resided during part of the study period. More precisely, for all cases we generate the starting times as in Section 2.4.1. Then for q_0 of the cases, one location is drawn uniformly in $\mathcal{C}$ and the remaining $E - 1$ locations are drawn uniformly in the unit square. Which location gets chosen in $\mathcal{C}$ is based on a function $G_{\mathcal{I}} : \mathbb{R}^+ \to [0, 1]$, the CDF of a lognormal distribution with median $m_{\mathcal{I}}$ and dispersion factor $\sigma_{\mathcal{I}}$. This function represents the distribution of the time length between exposure and diagnosis. Given the starting times, we use it to define the probability that exposure has occurred at each of the location:

$$p_k = \frac{G_{\mathcal{I}}(\tau - T_k) - G_{\mathcal{I}}(\tau - T_{k+1})}{G_{\mathcal{I}}(\tau)} \text{ for } k = 1, \ldots, E.$$

We then select one rank based on a multinomial distribution of size 1 and probabilities $(p_1, \dots, p_E)$ and assign the location with that rank to be drawn uniformly in $\mathcal{C}$. The starting times and shape of $G_{\mathcal{I}}$ determine at what period exposure most likely occurred. However the multinomial may select another rank.

2.5 Evaluation

We represent the unit square by a regular grid $\{y_1, y_2, \dots, y_r\}$ of $r = 50 \times 50 = 2500$ points. At each grid point, we implement our disease mapping function Γ from Section 2.3 using $N = 40$ and $p_0 = 0.1$, with one of the three weighting schemes described in Section 2.3.2 (Diagnosis, Sojourn, Incubation). This is performed for 1000 samples of $n = 100$ cases drawn from $F_{\tilde{X}}$. To evaluate our mapping method we use the metric proposed in Jeffery et al. [12, 13]. The residential history distribution of the cases is based on a spatial distribution that dichotomizes the risk of disease across the region. That is, q_0 of the cases are more likely to have partly resided in $\mathcal{C}$ than in the rest of the region. If a disease mapping function performs well, the disease scores $\Gamma(y_1), \dots, \Gamma(y_r)$ should have high values in $\mathcal{C}$ and low values outside of $\mathcal{C}$. Thus, given a threshold γ, the grid points belong to one of four categories: $\{y \in \mathcal{C}, \Gamma(y) > \gamma\}$, $\{y \notin \mathcal{C}, \Gamma(y) > \gamma\}$, $\{y \in \mathcal{C}, \Gamma(y) \leq \gamma\}$, $\{y \notin \mathcal{C}, \Gamma(y) \leq \gamma\}$. We [12, 13] define two metrics based on these categories, *sensitivity* and *specificity*. The former is the proportion of grid points in $\mathcal{C}$ that are given a 'high' disease score, while the latter is the proportion of grid points not in $\mathcal{C}$ that are given a 'low' disease score. To choose the threshold γ, we draw 100 samples of size 100 from a distribution similar to $F_{\tilde{X}}$, where, instead of fixing the radius of $\mathcal{C}$ at 0.1, we select it uniformly in (.05,.3). For each sample and a range of thresholds, we select the threshold that minimizes the sum $(1 - \mathit{sensitivity})^2 + (1 - \mathit{specificity})^2$, which corresponds to the distance between a point on the ROC curve and (0,1). Finally we define γ as the median of these 100 optimal thresholds.

3. Results

To apply our simulations, we choose some of the remaining parameters based on results published by Armenian and Lilienfeld [16]. We select the median and dispersion factor for leukemia after radiotherapy for ankylosing spondylitis, ($m_{\mathcal{I}} = 6.4$, $\sigma_{\mathcal{I}} = 1.71$) and fix $\tau = 20$ years. Also we choose the average time between locations as $1/\lambda = 4$ years. This value relates to a study published on duration of residence in the United States [4]. Finally we select several values for q_0: 10%, 25%, 35%, 50%, 75% and 100%.

3.1 One simulation

Figure 3 shows an application of DBM for residential history to one simulation when $q_0 = 50\%$. The color cutoffs are determined by resampling from the reference population F_0. The higher risk circle is identified more accurately with higher scores when mapping using Sojourn or Incubation weights rather than by mapping using only location at diagnosis (Diagnosis).

3.2 One thousand simulations

Figures 4 and 5 show the respective distribution of sensitivity and specificity from 1000 simulations. Sensitivity and specificity both tend to increase as the percentage

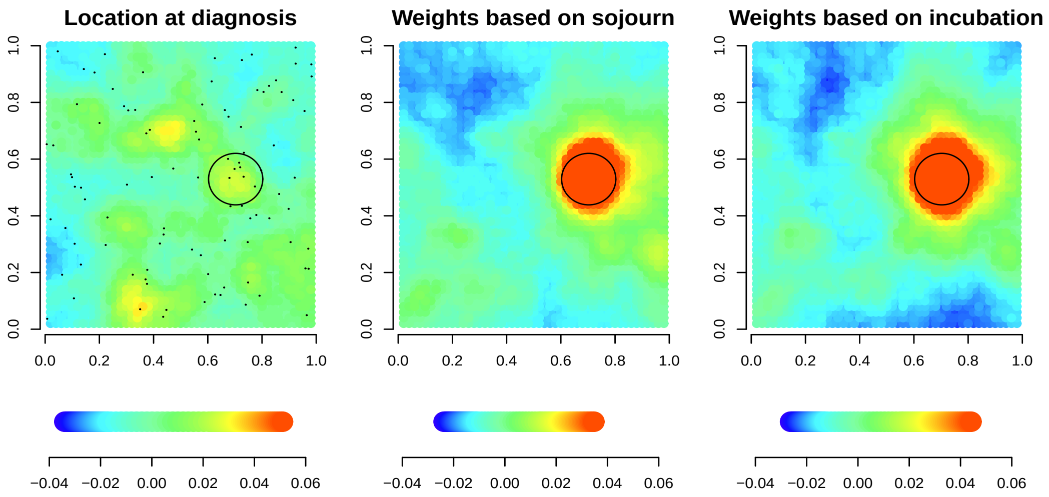

Figure 3: 50% of the cases have at least one location in the circle ($N = 40, p_0 = 0.1$). Each panel corresponds to one of the function G from Section 2.3.2. Additionally the left panel show the 100 locations at time of diagnosis.

of exposed cases increases. In terms of both measures, mapping with Sojourn weights improves the identification of the high risk circle compared to using only location at diagnosis, mostly when up to 50% cases are exposed. In fact the mapping with this weighting approach gives similar results to mapping with Incubation weights.

4. Conclusion

This work presents an extension of disease mapping which incorporates the residential histories of cases, adjusting each of the multiple locations by a weight. This new method allows us to explore association between disease status and past geographical exposure. These types of relationships are important to measure when particular areas of a region are linked to the disease. Simulations show that the accuracy in mapping a dichotomized risk in the unit square varies on whether or how the residential histories of individuals are taken into account by the weighting scheme. The method performs best when weights are based on accurate information about the time between exposure and diagnosis.

The weights are constructed as a finite partition of the interval $[0, 1]$ defined by a CDF G, and sum to one within an individual. Although the time information was used with two continuous example for G, one could also define it as a step function. The function G can also be chosen depending on which locations are most informative, recent ones or older ones. The incubation distribution of disease or an estimate of it is a reasonable choice, if available. One can imagine other weighting schemes if we relax the constraint that G is increasing. If no time information is available for example, all locations can be assigned the same weight ($G(u) = 1/E$). If the disease has short incubation period, mobility of individuals during the course of day can also act as a basis for weighting scheme, such as the proportion of time spent at each location (work/home). Alternatively, the weights could relate to covariates that vary from location to location, like age, and be adjusted to emphasize or deemphasize particular strata, in the case of diseases that affect individuals only during specific periods of their lifetime.

Future work should address some of the limitations present in this work, both

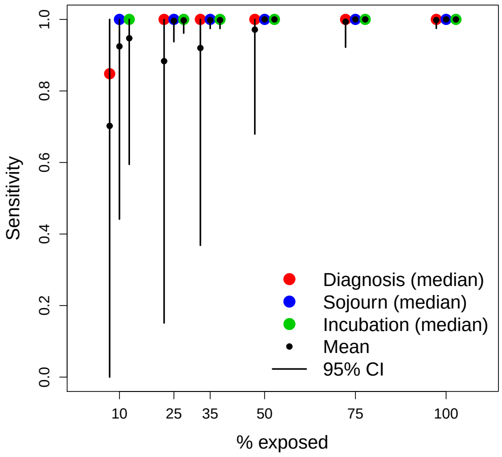

Figure 4: Sensitivity to locate high risk area (Y axis) for several percentage of cases exposed (X axis): median (colored dot), mean (black dot) and 95%CI (black line) according to weighting scheme (Red=Diagnosis, Blue=Sojourn, Green=Incubation), from 1000 simulations.

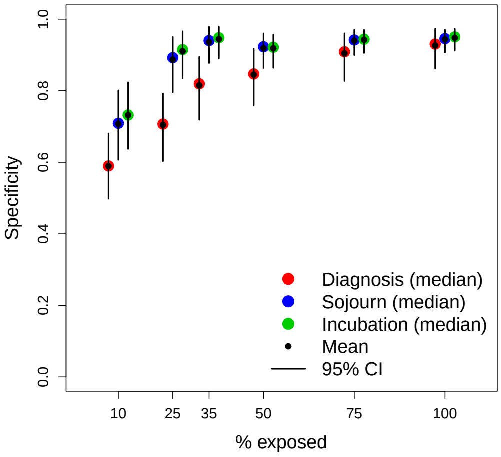

Figure 5: Specificity to locate high risk area (Y axis) for several percentage of cases exposed (X axis): median (colored dot), mean (black dot) and 95%CI (black line) according to weighting scheme (Red=Diagnosis, Blue=Sojourn, Green=Incubation), from 1000 simulations.

relative to the simulations and the assumptions made in the methods. The results of our simulations mostly show very high sensitivity regardless of the percentage of cases exposed, thus it is difficult to clearly distinguish the weighting schemes as well as for specificity. A large exposure radius might give a clearer contrast, although possibly at the expense of increased specificity.We also limited this study both to a uniform distribution in the unit square and an atemporal dichotomized risk. However, both aspects allow us to understand the methodology in a simple setting, and focus on the impact of a small number of already important parameters. The methodology presented in this work has been developed under the assumption that the reference population is known. In practice, it is estimated from a sample of individuals chosen to represent the null distribution. Aside from the difficulty in accessing a second dataset with residential history, the method so far does not take into account possible bias or variability occurring from this second selection. Finally, we have not explored how missing spatial information on individuals impacts our method. One possible approach is to redistribute the missing locations equally among the known ones, assuming the missingness pattern is ignorable [9, 10].

References

[1] A. Ozonoff, T. Webster, V. Vieira, J. Weinberg, D. Ozonoff, and A. Aschengrau. Cluster detection methods applied to the upper cape cod cancer data. *Environmental Health: A Global Access Science Source*, 4(1):19, 2005.

[2] L. Forsberg, M. Bonetti, C. Jeffery, A. Ozonoff, and M. Pagano. Distance based methods for spatial and spatio-temporal surveillance. *Spatial and Syndromic Surveillance for Public Health*, pages 133–152, Wiley(2005).

[3] L. Forsberg, C. Jeffery, A. Ozonoff, and M. Pagano. A spatiotemporal analysis of syndromic data for biosurveillance. *Statistical Methods for Counter-*

Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication, pages 173–193, Springer(2006).

[4] J. P. Schacter and J.J. Kuenzi. Seasonality of moves and the duration and tenure of residence: 1996. *United States Census Bureau*, 2002.

[5] V. Neumann, S. Günther, K.M. Müller, and M. Fischer. Malignant mesothelioma–german mesothelioma register 1987–1999. *International archives of occupational and environmental health*, 74(6):383–395, 2001.

[6] SW Lagakos, BJ Wessen, and M. Zelen. An analysis of contaminated well water and health effects in woburn, massachusetts. *Journal of the American Statistical Association*, pages 583–596, 1986.

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

[8] J. Cuzick and R. Edwards. Spatial clustering for inhomogeneous populations. *J Royal Statist Soc B*, 52:73–104, 1990.

[9] J. Manjourides. *Distance based methods for space time modelling of the health of populations.* PhD Thesis, Harvard University, Cambridge, MA, 2009.

[10] J. Manjourides and M. Pagano. Improving the power of chronic disease surveillance by incorporating residential history. *Submitted.*

[11] M. Bonetti and M. Pagano. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Statistics in Medicine*, 24(5):753–773, 2005.

[12] C. Jeffery. *Disease Mapping and Statistical Issues in Public Health Surveillance.* PhD Thesis, Harvard University, Cambridge, MA, 2010.

[13] C. Jeffery, A. Ozonoff, L.F. White, and M. Pagano. Locating spatial clusters in a surveillance setting. *Submitted.*

[14] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley-Interscience, 1992.

[15] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Chapman & Hall New York, 1995.

[16] H.K. Armenian and A.M. Lilienfeld. The distribution of incubation periods of neoplastic diseases. *American Journal of Epidemiology*, 99(2):92–100, 1974.