Spatial Risk Modeling for Infectious Disease Surveillance Using Population Movement Data

Roberto C.S.N.P. Souza[1], **Renato Assunção**[1], **Daniel B. Neill**[2], **Luís G.S. Silva**[3], **Wagner Meira Jr.**[1]
[1] Department of Computer Science, Universidade Federal de Minas Gerais
[2] Department of Computer Science & Center for Urban Science and Progress, New York University
[3] Department of Statistics, Universidade Federal de Minas Gerais
{nalon,assuncao,meira}@dcc.ufmg.br, daniel.neill@nyu.edu, lgsilva@ufmg.br

Abstract

We present three recently proposed subset scan methods for spatial disease surveillance that use movement data from case and control individuals, rather than a single location per individual, in order to identify areas with a high relative risk of infection. We illustrate the use of these methods to detect spatial clusters of dengue infection risk using geo-located data from Twitter classified into infected cases and non-infected controls.

1 Introduction

The spatial cluster detection task aims at detecting localized spatial regions or zones, called *spatial clusters*, where the probability of some event occurrence is higher than in the rest of the map. Spatial cluster detection methods, such as the spatial and subset scan statistics [Kulldorff, 1997, 2001, Neill, 2012], search the data to uncover the location and boundaries of any possible clusters. These methods usually work in an unsupervised manner, without prior knowledge of the relevant spatial patterns of anomalies such as their center, shape, or size. They also provide meaningful statistical measures to evaluate the significance of detected clusters.

The spatial scan statistic [Kulldorff and Nagarwalla, 1995, Kulldorff, 1997] is the most commonly used method in this class. It searches over a large set of geographical areas $\mathcal{Z}$ with a rigid circular shape, allowing the radius of each circle to vary. Over this set of regions, the spatial scan maximizes a likelihood ratio statistic, given by

$$L(\mathcal{Z}) = \frac{\mathbb{P}\left(\mathrm{Data} \mid \mathrm{H}_1(\mathcal{Z})\right)}{\mathbb{P}\left(\mathrm{Data} \mid \mathrm{H}_0\right)}\,, \tag{1}$$

where Data is specified according to the model, e.g., the distribution of disease cases over space and time. Across the several variants of the spatial scan statistics [Shi and Janeja, 2009, Tango and Takahashi, 2005, Neill et al., 2004], there has been one invariant aspect of the geographical characterization of the subjects: there is one and only one spatial position associated with each individual, whether that location represents a pixel in a medical image, as in Somanchi et al. [2018], or a spatially localized event, such as a crime or accident location. In particular, in health surveillance, these systems usually locate each individual by their home address and, more rarely, their workplace address. However, human mobility plays a key role in the transmission of infectious diseases [Stoddard et al., 2009], and relying solely on an individual's residential or workplace address as a proxy for the place that individual was infected ignores a multitude of exposures that individuals are subjected to during their daily routines.

In this paper, we explore different models to search for localized spatial risk clusters based on data from two groups of individuals, cases and controls. The cases are composed of individuals who

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

experienced a particular event related to the risk, such as crime victims or diseased individuals. The control group is composed of individuals who have not experienced that event during the same time period. Each individual is represented by a set of points describing their movement in space, specifying a location for each observation of that individual. Below we describe three probabilistic models that can be used in a spatial or subset scan statistic approach to search for the most likely spatial risk cluster. The models differ in their assumptions about the difference in risk between the case and control groups assuming that a spatial risk cluster exists. We illustrate the use of our models for discovery of spatial risk clusters of dengue, an important health problem in tropical areas. Geo-tagged Twitter data from two groups of individuals were analyzed. The individuals were classified into controls or disease cases based on the textual content of their messages. The methods were able to detect spatial clusters that are prime suspects for further epidemiological investigation.

2 Detecting Spatial Clusters in Mobility Patterns

We use Figure 1 to explain the problem. Each individual is indexed by an integer i and has a set of n_i spatial positions. These positions may come from lat-long coordinates of geo-located tweets, call detail records, or other sources. The positions from a single person are connected by line segments. Individuals are additionally labeled by two colors according to their status: cases (in red) or controls (in blue). In a disease surveillance scenario, the cases may represent the group of infected individuals and the controls are the non-infected population at risk. The positions marked with a hatched shadow are the ones which helped to identify the case individuals, for instance, when they mention specific dengue-related keywords in that tweet. It also shows a spatial zone $\mathcal{Z}$ where the risk of becoming a case given that an individual travels in that zone might be higher than in the rest of the region. Our main objective is to search for spatial clusters where the infection risk is significantly higher than elsewhere. A key challenge is that the number of positions n_i composing each mobility pattern can vary substantially between individuals i. Thus, simple approaches like counting the total numbers of case and control tweets per location would be biased and inaccurate; moreover, individuals with larger numbers of positions may be more likely to be identified as a case, since we have more information about them. Nevertheless, our assumption is that the entire mobility patterns will be informative of the riskier areas if we properly compare the spatial patterns from case and control individuals. The problem is to find appropriate ways to make this pattern comparison. The multiple locations associated with each individual, rather than the usual single location (such as their place of residence), lead us to consider several different models, which we describe next.

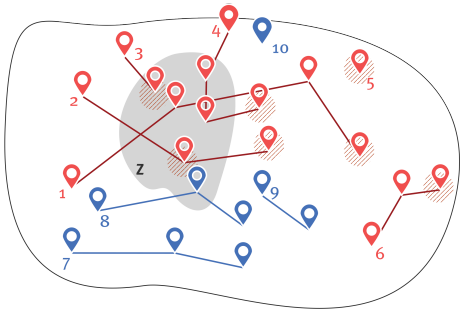

Figure 1: Schematic drawing of the problem.

2.1 Models

In this section, we briefly present three spatial models that deal differently with the data: the previously proposed visit and infection models [Souza et al., 2016], and our new logistic exposure model [Souza et al., 2018]. To make it more concrete, we describe the models in the context of data coming from geo-tagged Twitter data. Intuitively, the visit model searches for the most likely zone $\mathcal{Z}$ looking at the conditional probability $\mathbb{P}$(i tweets from $\mathcal{Z}$ | i is a case) while the infection model considers the inverse conditional probability $\mathbb{P}$(i is a case | i tweets from $\mathcal{Z}$). The third model conditions on the total number of points n_i associated with the i-th individual.

Visit Model. Let $V_{i,z}$ be the number of tweets issued in $\mathcal{Z}$ among the n_i total number of tweets the i-th individual generated. Let $\mathbb{1}[V_{i,z} \geq 1]$ be the binary random variable indicating whether the i-th individual ever tweeted inside the candidate zone $\mathcal{Z}$. These random variables can be assumed independent but they are not identically distributed, as the success probability depends on the number n_i of tweets issued by each individual. Denote by $p = p(\mathcal{Z})$ the probability that, given that a case individual is tweeting, he does so from within $\mathcal{Z}$. Let $\bar{p} = \bar{p}(\mathcal{Z})$ be the similar probability for

a control individual. We are interested in zones where $p(Z) > \bar{p}(\mathcal{Z})$. For a case user, we have $\mathbb{P}(V_{i,z} \geq 1) = 1 - (1-p)^{n_i}$ and, for a control subject, it is equal to $1 - (1-\bar{p})^{n_i}$.

Infection Model. We will estimate the probability that an individual becomes a case given that they visited region $\mathcal{Z}$ a total of k times. Let $r = r(\mathcal{Z})$ be the infection risk inside the candidate cluster and $\bar{r} = r(\bar{\mathcal{Z}})$ the infection risk in $\bar{\mathcal{Z}}$, the region outside $\mathcal{Z}$. We are interested in zones $\mathcal{Z}$ where $r(\mathcal{Z}) > r(\bar{\mathcal{Z}})$. Let I_i be the binary indicator that the i-individual is a case. We assume that these binary random variables are independent. They are not identically distributed since the probability of $I_i = 1$ depends on the number of visits $V_{i,z}$ made by the i-th individual to the zone $\mathcal{Z}$. We have $\mathbb{P}(I_i = 1 | V_{i,z} = k_i) = 1 - \mathbb{P}(I_i = 0 | V_{i,z} = k) = 1 - (1-r)^{k_i} (1-\bar{r})^{n_i - k_i}$.

Logistic Model. Let Y_i be the binary indicator that the i-th individual is a case rather than a control. The probability that someone appears as a case increases with the number of points n_i in which we have information. We allow for this effect through a possibly non-linear, monotone non-decreasing function $g(n_i)$:

$$g(n_i) = \frac{\mathbb{P}(Y_i = 1 | n_i)}{\mathbb{P}(Y_i = 0 | n_i)}, \tag{2}$$

where $g(n_i)$ is an arbitrary and unspecified function. The proportion $p(\mathcal{Z})_i$ of time spent on the putative high risk zone $\mathcal{Z}$ modifies this ratio according to the ratio $\lambda_{\text{in}}/\lambda_{\text{out}}$ between the risk inside and outside $\mathcal{Z}$:

$$\frac{\mathbb{P}(Y_i = 1 | n_i, p(\mathcal{Z})_i)}{\mathbb{P}(Y_i = 0 | n_i, p(\mathcal{Z})_i)} = \frac{\mathbb{P}(Y_i = 1 | n_i)}{\mathbb{P}(Y_i = 0 | n_i)} \left(\frac{\lambda_{\text{in}}}{\lambda_{\text{out}}} \right)^{(p(\mathcal{Z})_i - p_0(\mathcal{Z}))} = g(n_i)\, e^{\beta\, (p(\mathcal{Z})_i - p_0(\mathcal{Z}))} \tag{3}$$

where $\beta = \log(\lambda_{\text{in}}/\lambda_{\text{out}})$. The term $p_0(\mathcal{Z}) = \mathbb{E}(p(\mathcal{Z})_i)$ is the expected value of the proportion $p(\mathcal{Z})_i$ over all individuals. When $\mathcal{Z}$ is indeed a high risk zone, we have $\beta > 0$ and, as a consequence, individuals spending a considerable proportion of their time inside $\mathcal{Z}$ (as estimated from the set of observed locations) have a larger probability of becoming a disease case. Hence, intuitively, zones where this β coefficient is large are candidate high risk zones. Model (3) implies a binomial distribution for Y_i with a semi-parametric logistic probability specification with the $g(n_i)$ as an offset:

$$\mathbb{P}(Y_i = 1 | n_i, p(Z)_i) = \frac{g(n_i)}{g(n_i) + \exp(-\beta(p(\mathcal{Z})_i - p_0(\mathcal{Z})))} \tag{4}$$

2.2 Inference

Due to space constraints, we derive the inference step only for the last model. Analogous reasoning can be carried out for the other models. We want to test the null hypothesis $H_0 : \lambda(x, y) = \lambda_{\text{all}}$ is constant versus the set of alternative hypotheses $H_1(\mathcal{Z})$: there is a region $\mathcal{Z}$ such that $\lambda(x, y) = \lambda_{\text{in}}$ for all $(x, y) \in \mathcal{Z}$ and $\lambda(x, y) = \lambda_{\text{out}} < \lambda_{\text{in}}$ for all $(x, y) \notin \mathcal{Z}$. This alternative hypothesis is equivalent to having $\beta > 0$ in model (4). The aim is to find the most likely zone $\mathcal{Z}$ given the evidence provided by $\mathbf{S}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{ik})$ and the spatial locations of the tweets. For model (4), considering a fixed region $\mathcal{Z}$, the likelihood for the observed sample $Y_1, Y_2, \ldots, Y_n$ of binary variables is given by the logistic likelihood

$$L(H_1, \mathcal{Z}, \beta) = \prod_i \mathbb{P}(Y_i = 1 | n_i, p(\mathcal{Z})_i)^{y_i} \mathbb{P}(Y_i = 0 | n_i, p(\mathcal{Z})_i)^{1 - y_i} \tag{5}$$

For fixed $\mathcal{Z}$, the maximum likelihood estimator of β maximizes (5) and it is denoted by $\hat{\beta}(\mathcal{Z})$. The most likely zone $\hat{\mathcal{Z}}$ is finally given by $\hat{\mathcal{Z}} = \arg\max_{\mathcal{Z}} L(H_1, \mathcal{Z}, \hat{\beta}(\mathcal{Z}))$. To obtain the p-value, it is useful to denote this most likely zone obtained with the observed dataset as $\hat{\mathcal{Z}}^{(0)}$. Under the null hypothesis $H_0 : \lambda_{\text{in}} = \lambda_{\text{out}} = \lambda$, we have $\beta = 0$ as staying longer in $\mathcal{Z}$ has no effect on the probability of $Y_i = 1$. Therefore, in the case of the logistic model, $L(H_0) = \prod_i \mathbb{P}(Y_i = 1 \mid n_i)^{y_i} \mathbb{P}(Y_i = 0 \mid n_i)^{1 - y_i}$ where $\mathbb{P}(Y_i = 1 \mid n_i) = 1/(1 + \exp(g(n_i)))$ as $\beta = 0$ under H_0.

To evaluate the statistical significance of the maximum likelihood estimator $\hat{\beta}(\mathcal{Z})$ obtained from this model, we calculate the maximum likelihood ratio test statistic (MLRT):

$$T^0 = \frac{L(H_1, \hat{\mathcal{Z}}, \hat{\beta}(\hat{\mathcal{Z}}))}{L(H_0)}.$$

Next, we run a permutation test to obtain its associated p-value. We randomly permute the case and control labels (i.e., randomly permute the observed values of Y_i) among the individuals. This guarantees that, in this permuted dataset, the cases and controls gain their labels in a manner disassociated with any spatial aspect. This permutation assignment is carried out a large number *nsim* of times. After the random assignments, we run the entire zone detection procedure with the pseudo datasets obtaining $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(nsim)}$, the associated most likely zones $\hat{\mathcal{Z}}^{(1)}, \ldots, \hat{\mathcal{Z}}^{(nsim)}$ and the value of the MLRT $T^1, T^2, \ldots, T^{nsim}$. The p-value is essentially the proportion of permutation-based values $T^1, T^2, \ldots, T^{nsim}$ that are larger than the observed value T^0.

3 Case Study: Dengue in Brazil

In this case study, we are interested in searching for spatial clusters associated to high risk of infection by dengue. Traditional systems for dengue surveillance place the case individuals at their home locations. Though easy to obtain, residential addresses are often a poor indicator of the regions where people and infected mosquitoes tend to interact more. Our geolocated data were collected through the Twitter Streaming API[1]. From a large number of analysis based on Brazilian municipalities, we selected the results from Sorocaba city, located in the Southeast region of Brazil, to illustrate our methods. We identify "infected" individuals (cases) as those individuals who have at least one tweet classified as a current, personal experience with dengue. This classification was made through NLP techniques applied to textual content of the tweets. Because of the incubation period and recovery time, infected Twitter users are likely to mention dengue in their tweets days after they are infected, and usually not at the location where the exposure (mosquito bite) occurred, which makes the task harder. The control individuals are composed of the remaining users. To run our models, we create the mobility patterns of each individual by retrieving the positions of all messages they issued in the period of analysis.

To run the models, we scan over axis-aligned rectangular regions of different sizes. The number of Monte Carlo simulations was set to 299 and the significance level $\alpha = 0.05$. The offset $g(n_i)$ was estimated using a Lowess smoother. Both Infection and Logistic models were able to detect at least one region in the selected city. The visit model only suspected regions but they were not significant at level $\alpha = 0.05$. As each model conditions its probability on different events they may identify different regions. This effect can be useful depending on the type of event being monitored. Figure 2 shows the prime suspect regions pinpointed by the logistic model. It is worth mentioning that, there are many points of interest, such as hospitals, parks, and college campus, inside the detected regions. As those places are non-residential, standard approaches would fail to consider them as potential infection places in the event of a spike in the number of cases.

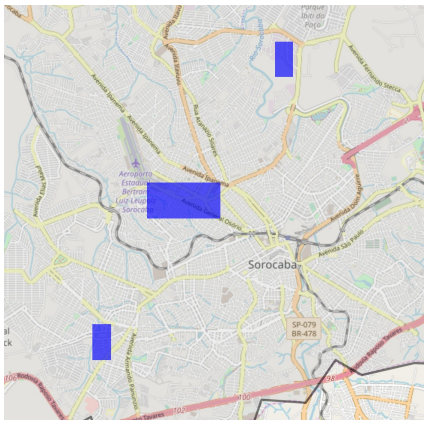

Figure 2: Zoom in to the regions detected by the Logistic Model.

4 Discussion and Concluding Remarks

Identifying places where people have higher risk of being infected, rather than focusing on residential address locations, may be key to guide spatial surveillance actions, specially for vector-borne diseases such as dengue, allowing public health officials to focus mitigation actions. The stochasticity of location data is not appropriate for typical spatial cluster detection tools such as the traditional spatial scan statistic [Kulldorff, 1997]. Each user is represented by a different number of geographic points and the variability of these numbers is large; traditional approaches can be easily misled if not extended to account for this special structure. We expect that our methods will also be useful to other spatial surveillance problems where movement data can bring relevant information.

[1] https://dev.twitter.com/streaming/overview

Acknowledgments

All the authors would like to thank FAPEMIG, CNPq and CAPES for their financial support. This work was also partially funded by projects InWeb (MCT/CNPq 573871/2008-6), MASWeb (FAPEMIG-PRONEX APQ-01400-14), EUBra-BIGSEA (H2020-EU.2.1.1 690116, Brazil/MCTI/RNP GA-000650/04), INCT-Cyber, (CNPq 465714/2014-5), ATMOSPHERE (H2020 777154 and MCTIC/RNP 51119) and by the Google Research Awards for Latin America program.

References

M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14(8): 799–810, 1995.

Martin Kulldorff. A spatial scan statistic. *Comm. in Stat. - Theory and Meth.*, 26(6):1481–1496, 1997.

Martin Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):61–72, 2001.

Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

Daniel B. Neill, Andrew W. Moore, Francisco Pereira, and Tom M. Mitchell. Detecting significant multidimensional spatial clusters. In *Advances in Neural Information Processing Systems*, pages 969–976, 2004.

Lei Shi and Vandana Pursnani Janeja. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). In *Proc. of the 15th SIGKDD*, pages 767–776, 2009.

Sriram Somanchi, Daniel B. Neill, and Anil V. Parwani. Discovering anomalous patterns in large digital pathology images. *Statistics in Medicine, in press*, 2018.

Roberto C. S. N. P. Souza, Renato Assunção, Derick Oliveira, Denise Brito, and Wagner Meira Jr. Infection hot spot mining from social media trajectories. In *Proc. of the ECML/PKDD*, 2016.

Roberto C. S. N. P. Souza, Renato Assunção, Daniel B. Neill, and Wagner Meira Jr. Subset scan methods for infection spatial cluster detection in social media mobility patterns. *working paper*, 2018.

Steven Stoddard et al. The role of human movement in the transmission of vector-borne pathogens. *PLoS neglected tropical diseases*, 3(7):e481, 2009.

Toshiro Tango and Kunihiko Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(1), 2005.