Detecting Spatial Clusters of Disease Infection Risk Using Sparsely Sampled Social Media Mobility Patterns

Roberto C.S.N.P. Souza*
Department of Computer Science
Universidade Federal de Minas Gerais
nalon@dcc.ufmg.br

Renato M. Assunção
Department of Computer Science
Universidade Federal de Minas Gerais
assuncao@dcc.ufmg.br

Daniel B. Neill
Center for Urban Science and Progress
New York University
daniel.neill@nyu.edu

Wagner Meira Jr.
Department of Computer Science
Universidade Federal de Minas Gerais
meira@dcc.ufmg.br

ABSTRACT

Standard spatial cluster detection methods used in public health surveillance assign each disease case to a single location (typically, the patient's home address), aggregate locations to small areas, and monitor the number of cases in each area over time. However, such methods cannot detect clusters of disease resulting from visits to non-residential locations, such as a park or a university campus. Thus we develop two new spatial scan methods, the unconditional and conditional spatial logistic models, to search for spatial clusters of increased infection risk. We use mobility data from two sets of individuals, disease cases and healthy individuals, where each individual is represented by a sparse sample of geographical locations (e.g., from geo-tagged social media data). The methods account for the multiple, varying number of spatial locations observed per individual, either by non-parametric estimation of the odds of being a case, or by matching case and control individuals with similar numbers of observed locations. Applying our methods to synthetic and real-world scenarios, we demonstrate robust performance on detecting spatial clusters of infection risk from mobility data, outperforming competing baselines.

CCS CONCEPTS

• **Information systems → Data mining**.

KEYWORDS

spatial scan statistics, social media data, spatial cluster detection

ACM Reference Format:
Roberto C.S.N.P. Souza, Renato M. Assunção, Daniel B. Neill, and Wagner Meira Jr.. 2019. Detecting Spatial Clusters of Disease Infection Risk Using Sparsely Sampled Social Media Mobility Patterns. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3347146.3359369

1 INTRODUCTION

Infectious and parasitic diseases account for a major part of the world disease burden. According to the World Health Organization, they are responsible for approximately 19% of all deaths worldwide but reaching more than 50% in poor countries [21]. Disease surveillance is a key public health approach to control and prevent the spread of infectious diseases, and spatial cluster detection (SCD) is a major component of disease surveillance systems. The primary goal of spatial cluster detection is to find a set of localized regions, named *spatial clusters*, where a certain event of interest has a higher probability of occurring than in the rest of the map. A major application of SCD in epidemiology is the detection of disease clusters to suggest risk factors, to focus preventive efforts, and for outbreak monitoring [14]. As the algorithms proposed for SCD are based on statistical data, there is always some uncertainty associated with detected clusters. Hence, methods in this class also provide meaningful statistical measures to evaluate whether there is enough evidence to call the detected regions a true spatial cluster or if they are likely to have appeared due to chance.

Standard methods for SCD, such as the spatial and subset scan statistics [12, 18] and their many variants, locate each individual case by home address and monitor the number of cases in each possible area over time. Depending on the application scenario, this approach may or may not extract enough relevant spatial information to enable effective disease surveillance. Determinants of cancer risk may vary on a coarse-grained spatial and temporal scale, and thus the residence address may be sufficient for detection. However, relying solely on home address may be inadequate for infectious diseases. For instance, human mobility plays a key role in mosquito-borne disease transmission such as dengue, yellow fever, and Zika [28], since people may be exposed to disease in any of the places where they spend their time. Therefore, identifying high-risk areas for such diseases requires richer geographical information than simply a single location per individual. Relying solely on residential address as a proxy for the place of infection in such cases ignores a multitude of exposures that individuals are subjected to during daily events. This simplification provides little information about the actual places where people are most likely to be infected.

[*] Part of this work was done while Roberto was a visiting scholar at New York University.

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/3347146.3359369

To date, the main difficulty in exploiting people's movements to search for areas of high infection risk has been the cost and time in obtaining such information, usually done through expensive and time consuming surveys. However, over the last decade, the ubiquity and pervasiveness of technology combined with a rapid increase in the number of mobile devices has enabled the large-scale generation, collection and storage of human geographical movement data. For instance, social media such as Twitter can provide rich and useful location information. Based on the textual content of messages we can identify diseased individuals and compare their mobility patterns to others. This data can bring essential information to public health by identifying places of increased infection risk, informing disease prevention and mitigation strategies.

In this paper, we develop new methods for accurate detection of spatial clusters associated with infectious diseases based on sparsely sampled mobility data of diseased and non-diseased individuals, named *cases* and *controls*, respectively. As the continuous spatial tracking of a large sample of infected and non-infected individuals would raise serious privacy issues, we instead analyze geo-located Twitter data (tweets), which are readily and publicly available but provide only occasional snapshots of each individual's movements. The key aspect in our methods is that the input for each individual is a series of locations, which we call mobility patterns, rather than a single location, such as the residence address. The number n_i of positions composing each mobility pattern can vary substantially between the individuals, indexed by i. As we show, simple approaches like counting the total numbers of case and control tweets per location are biased and inaccurate. Thus we propose two new spatial scan methods, the *unconditional* and the *conditional* spatial logistic models, which account for the multiple, varying number of spatial locations per individual. Both models use the proportion of an individual's tweets in each location as an estimate of the proportion of time spent in that location. This estimate is biased as the individuals' propensity to tweet is not spatially constant. However, this bias is cancelled by our use of a control sample. Our unconditional model models the variable contribution of each individual through direct estimation of the odds of being a case using a semi-parametric logistic specification. As an alternative that does not require this estimation, we propose a case-control matching strategy in the conditional model to control for the number of tweets. We search for localized regions where the infection risk is substantially higher than in the rest of the map by maximizing a log-likelihood ratio scan statistic, thus providing a non-trivial generalization of the traditional spatial scan to mobility data with multiple locations per individual. We perform an extensive experimental study applying our methods to both synthetic and real-world scenarios, demonstrating robust performance on detecting spatial clusters from sparsely sampled mobility data.

1.1 Motivating Scenario

This work has been motivated by an important public health problem in tropical countries: vector-borne diseases. Despite all the surveillance actions and countermeasures, such diseases still challenge health services and policy makers, particularly in developing countries. For instance, dengue is regarded as the most important mosquito-borne viral disease [2, 17]. The World Health Organization (WHO)[1] estimates that almost half of the world's population is at risk of infection with dengue viruses, concentrated in South and Central America, Asia and Pacific regions [2, 17]. There is no currently approved, effective and broadly available vaccine to protect the population against the virus. Preventing dengue depends entirely on controlling the mosquito vectors or interruption of human–vector contact [22]. Therefore, epidemiological surveillance and effective vector control are the mainstay of dengue prevention, although recent studies have questioned the correlation between vector prevalence and dengue transmission [6].

As described above, traditional surveillance systems for dengue place diseased individuals at their home address, since residence information was historically the only available geographic data, and monitor the increase in the number of cases at each location. On the other hand, human mobility plays a key role in dengue transmission, especially given that the mosquitoes which spread dengue are most active during the day [28]. Therefore, residence location may be a poor indicator of the actual regions where humans and infected vectors tend to interact more. Being able to identify such high-risk areas would greatly benefit disease surveillance for dengue and similar vector-borne infectious diseases by targeting preventive efforts and mitigation actions where they are most needed.

2 BACKGROUND

The spatial scan statistic [12] is a powerful and widely used method for spatial cluster detection. Let N be the number of individuals in a map, with C of them being disease cases. Each individual is located in a single position in the map. Let Z be an arbitrary region. Under the alternative hypothesis $H_1(Z)$ that Z is an area of increased disease risk, the Bernoulli spatial scan statistic [12] assumes that, for all individuals, the probability of being a case in Z is p, while outside Z this probability is q, $p > q$. Let n_z and c_z be the total number of individuals and the number of cases individuals in Z, respectively. The likelihood function for the Bernoulli model is given by:

$$L(Z, p, q) = p^{c_Z}(1-p)^{n_Z - c_Z} q^{C - c_Z}(1-q)^{(N-n_Z)-(C-c_Z)} .$$

The method then 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 log-likelihood ratio statistic given by:

$$LLR(Z) = \log \frac{\mathbb{P}(\text{Data} \mid H_1(Z))}{\mathbb{P}(\text{Data} \mid H_0)} . \quad (1)$$

The null hypothesis H_0 assumes complete spatial randomness, i.e., each individual is equally likely to be a case everywhere in the map and thus $p = q$. After maximizing Equation (1) over all considered circular regions to identify the most likely cluster, the method computes the statistical significance of the detected cluster through Monte Carlo hypothesis testing.

The development of the spatial scan statistics opened the door for many additional research directions in spatial cluster detection, including: (i) overcoming the limitation of a rigid circular scanning window by allowing elongated [20], elliptical [13], linear [23] and

[1] http://www.who.int/denguecontrol

irregularly-shaped regions [1, 5, 7, 29]; (ii) reducing the computational efforts in the search for anomalous regions [15, 18, 31]; (iii) considering different representations with other parametric and also non-parametric models [3, 9]; (iv) expanding application scenarios aside from disease surveillance, for instance, by targeting identification of hot spots zones associated to crime events and traffic accidents [19, 23]; and (v) considering different data types such as categorical, graph and image [3, 16, 24].

A common premise in almost all SCD methods is that there is only a single spatial position associated with each individual, sometimes the exact lat/long coordinates (e.g., corresponding to home or work address), sometimes an area (e.g., census tract) in which an event occurred, or sometimes a pixel in an image [24]. There are few studies where more than one location associated with an individual is considered. For example, as the time between initial exposure to a carcinogen and cancer diagnosis can take decades, Jacquez et al. [10, 11] took into account the previous home locations of individuals to infer spatial determinants of cancer risk. However, Jacquez et al. require that each individual's location and their k-nearest neighbors are (i) infrequently changing and (ii) known at all times. This makes sense for residential histories, but not for daily movements, where locations change continuously and are only intermittently observed. The number of locations associated with each individual in [10, 11] is quite small, not exceeding two for most individuals.

When searching for places where people are more likely to be infected by a given disease, environmental exposures can play a fundamental role. In these cases, access to reasonably fine-grained mobility traces can bring significant information about such exposures and therefore about actual places of infection. In this direction, [26] exploited readily available social media data to track infected and uninfected individuals aiming at uncovering infection hotspots. Their methods substantially improve on the spatial scan statistics in different scenarios using mobility data [25]. However, their simplifying assumptions can lead to highly undesirable properties. For instance, they assume that tweets are independent and drawn from the same location distribution across all individuals. This way, an individual with many tweets would be assumed very likely to visit every location at least once. Also, people with many tweets are almost certain to be infected according to their model. These are severe data modeling shortcomings that must be overcome. In the next section, we introduce a new statistical framework that takes into account all these aspects to provide a more realistic data generating model, and our experiments in Section 4 demonstrate large performance improvements compared to these previous models.

3 DETECTING INFECTION CLUSTERS WITH SOCIAL MEDIA DATA

We are interested in monitoring a certain risk population in a given geographical region $\mathcal{R}$ for infection events by a certain disease. Each event is formally represented by a three-dimensional point specified by its spatial and temporal coordinates. An individual moves around in space and his position at time t is given by $(x(t), y(t))$. Figure 1 shows the spatial trajectory $\mathcal{T}$ of a given individual during a certain time period. Let $B \subset \mathcal{R}$ be a sub-region and $N(B, [t_1, t_2))$ be the number of events within B in the time period $[t_1, t_2)$. We are interested in locating the geographic regions where the risk of getting infected is higher than elsewhere. These are shown as the regions Z_1 and Z_2 in Figure 1.

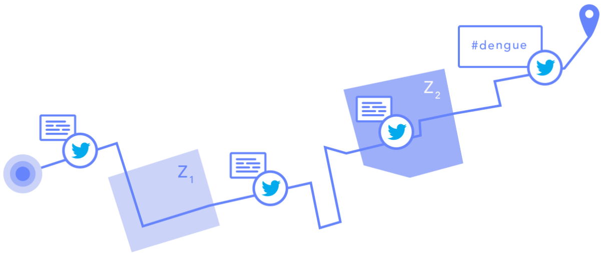

Figure 1: Schematic illustration of the spatial trajectory of a given diseased individual as well as his tweeting behavior.

Point processes are a stochastic framework developed to model discrete random events; a useful and flexible model in this class is the inhomogeneous Poisson point process. Let $\lambda(x, y) \geq 0$ be the intensity function at the geographic position (x, y) in a given time unit. The surface $\lambda(x, y)$ is defined on the entire region of interest $\mathcal{R}$ and its height at (x, y) is proportional to the infection risk for an individual exposed at this location. A more intuitive interpretation of this intensity function is given by the expected number of infection events after an individual follows a given path $\mathcal{T}$ such as that shown in Figure 1. Let the path be parameterized by the time so $\mathcal{T} = \{(x(t), y(t)), t \in [t_1, t_2)\}$ is the trajectory traveled in the time period $[t_1, t_2)$. We represent the length of the trajectory $\mathcal{T}$ by $\ell(\mathcal{T})$. The expected number of disease episodes for someone exposed to the trajectory $\mathcal{T}$ during the observation period is the line integral of $\lambda(x, y)$ over $\mathcal{T}$:

$$\mathbb{E}\left(N(\mathcal{T})\right) = \int_{t_1}^{t_2} \lambda(x(t), y(t))\, dt \; . \tag{2}$$

We partition the observation region into $\mathcal{R} = \mathcal{Z} + \mathcal{Z}^c$ where $\mathcal{Z}$ is a not necessarily connected sub-region where the intensity function is higher than a threshold. In Figure 1, for example, $\mathcal{Z} = Z_1 \cup Z_2$. We make the simplifying assumption that $\lambda(x, y) = \lambda_{\text{in}}$ for $(x, y) \in \mathcal{Z}$ and $\lambda(x, y) = \lambda_{\text{out}}$ for $(x, y) \in \mathcal{Z}^c$ with $\lambda_{\text{in}} > \lambda_{\text{out}}$. In this case, the line integral (2) can be written as

$$\begin{aligned} \mathbb{E}\left(N(\mathcal{T})\right) &= \lambda_{\text{in}} \int_{\mathcal{T} \cap \mathcal{Z}} dt + \lambda_{\text{out}} \int_{\mathcal{T} \cap \mathcal{Z}^c} dt \\ &= \lambda_{\text{in}}\, \ell(\mathcal{T} \cap \mathcal{Z}) + \lambda_{\text{out}}\, \ell(\mathcal{T} \cap \mathcal{Z}^c) \\ &= \ell(\mathcal{T}) \left(\lambda_{\text{out}} + (\lambda_{\text{in}} - \lambda_{\text{out}}) \frac{\ell(\mathcal{T} \cap \mathcal{Z})}{\ell(\mathcal{T})} \right) \\ &= \ell(\mathcal{T}) \left(\alpha + \beta p(\mathcal{Z}) \right) \; . \end{aligned} \tag{3}$$

where $\ell(\mathcal{T} \cap \mathcal{Z})$ is the length of the path $\mathcal{T}$ that lies within $\mathcal{Z}$ and $p(\mathcal{Z})$ is the proportion of time spent in $\mathcal{Z}$. Hence, the expected number of disease episodes increases linearly with the exposure time $\ell(\mathcal{T})$ and with the relative amount of time spent in the high risk zone $\mathcal{Z}$. The β coefficient measures the absolute risk increase of $\mathcal{Z}$ with respect to the baseline $\mathcal{Z}^c$.

Let D be a binary random variable indicating that the individual had at least one disease episode during his exposure period. Then,

$$\begin{aligned}\mathbb{P}(D=1) &= 1-\mathbb{P}(D=0) = 1-\exp\left\{-\int_{t_1}^{t_2}\lambda(x(u),y(u))du\right\} \\ &= 1-\exp\left\{-\ell(\mathcal{T})\left(\alpha+\beta p(\mathcal{Z})\right)\right\} \qquad (4)\end{aligned}$$

3.1 Social Media Data and Disease Infection

The main difficulty in learning model (4) from data is that we do not have access to the continuous trajectory $\mathcal{T}$ of an individual but rather only sparse and relatively few spatial positions along his path in certain time moments. We will use our motivating scenario – dengue – to illustrate the problem. In Figure 1, for example, these finite number of positions are given by the geo-tagged Twitter posts marked by Twitter birds with their associated text. Each tweet is classified as a *positive* or as *negative* tweet based on its content.

In addition to this difficulty, we also do not have the true health status D of the individuals. Instead, we are left only with the tweets at each spatial position from which we need to infer the health status. As dengue is a major debilitating disease, we can expect that sick or recovering individuals are likely to mention this fact if they ever engage on social media around the disease episode. This is the reason why we see many tweets with explicit mention to dengue occurrence at an individual level. Tweets mentioning personal experience with dengue are labeled as positive tweets; otherwise, they are labeled as negative tweets. For instance, in Figure 1, only one of the tweets is positive among the five issued by that individual.

We connect the tweets information with the epidemiological model from Section 3. The i-th individual has n_i tweets posted at times $t_1, t_2, \ldots, t_{n_i}$. Let $W(t_j)$ be the binary indicator that a tweet posted at time t_j is a positive or negative tweet. We define a binary variable Y_i as $Y_i = 0 \iff \left[W(t_1)=0,\ldots,W(t_{n_i})=0\right]$. That is, Y_i is a binary variable indicating whether the i-th individual ever tweeted about dengue disease. The binary variables D_i and Y_i are not equal but they are strongly correlated. Naturally, we expect that $\mathbb{P}(D_i=1|Y_i=1) > \mathbb{P}(D_i=0|Y_i=1)$: individuals tweeting about a personal experience with dengue are more likely than not to have been diseased. We also can expect that $\mathbb{P}(Y_i=1|D_i=1) > \mathbb{P}(Y_i=1|D_i=0)$ because $\mathbb{P}(D_i=1|Y_i=1) > \mathbb{P}(D_i=0|Y_i=1) \iff \mathbb{P}(Y_i=1|D_i=1) > \mathbb{P}(Y_i=1|D_i=0)\frac{\mathbb{P}(D_i=0)}{\mathbb{P}(D_i=1)}$, and $\mathbb{P}(D_i=0)$ is much larger than $\mathbb{P}(D_i=1)$: in our dataset, only 0.67% of individuals were labeled with $D_i=1$. That is, diseased people (with $D_i=1$) are more likely to tweet about dengue as a personal experience and hence to have $Y_i=1$.

The probability $\mathbb{P}(Y_i=1)$ depends on the number n_i of tweets in an undesirable way. Everything else being equal, an individual with only a handful of tweets during one entire year has less chance of appearing as a positive $Y_i=1$ case than another individual with hundreds of tweets. Typically, the tweets of the second individual constitute a better report of what happens to him. If he ever gets dengue, he will more likely tweet about it. The first individual tweets only very occasionally and hence most of his life goes unobserved in the Twitter data, including a possible disease episode.

We propose two models allowing for the influence of both the number n_i of tweets and the proportion $p(\mathcal{Z})_i$ of time that individual i spends in region $\mathcal{Z}$ on the probability $\mathbb{P}(Y_i=1)$. The first model requires direct estimation of the effect of n_i on Y_i. The second model is able to eliminate this confounding factor by matching case and control individuals based on their number of tweets.

3.2 Model 1: The Unconditional Spatial Logistic Model

Given the binary nature of the variable Y_i, we adopt a logistic formulation where n_i enters through a possibly non-linear monotone non-decreasing function $g(n_i)$. The effect of n_i on the probability of being a positive case is modeled through the odds

$$g(n_i) = \frac{\mathbb{P}(Y_i=1|n_i)}{\mathbb{P}(Y_i=0|n_i)} \qquad (5)$$

where the function $g(n_i)$ is fit by semi-parametric estimation. To do so, we split users' n_i into ranges and compute the odds of being a case in each range. Then, we use either a linear model (if there is a power-law relationship between n_i ranges and the odds) or locally-weighted linear regression (loess) otherwise. The proportion $p(\mathcal{Z})_i$ of time spent on the putative high risk region $\mathcal{Z}$ modifies this ratio according to the ratio between the risk inside and outside $\mathcal{Z}$:

$$\begin{aligned}\frac{\mathbb{P}(Y_i=1|n_i,p(\mathcal{Z})_i)}{\mathbb{P}(Y_i=0|n_i,p(\mathcal{Z})_i)} &= \frac{\mathbb{P}(Y_i=1|n_i)}{\mathbb{P}(Y_i=0|n_i)}\left(\frac{\lambda_{\text{in}}}{\lambda_{\text{out}}}\right)^{(p(\mathcal{Z})_i-p_0(\mathcal{Z}))} \\ &= g(n_i)\, e^{\beta\ (p(\mathcal{Z})_i-p_0(\mathcal{Z}))} \qquad (6)\end{aligned}$$

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 region, we have $\beta > 0$ and, as a consequence, individuals spending a considerable proportion of their time inside $\mathcal{Z}$ have an increased probability of becoming a disease case.

Model (6) implies a binomial distribution for Y_i with a semi-parametric logistic probability specification:

$$\mathbb{P}(Y_i=1|n_i,p(Z)_i) = \frac{g(n_i)}{g(n_i)+\exp(-\beta(p(\mathcal{Z})_i-p_0(\mathcal{Z})))} \qquad (7)$$

While model (4) gives support to the presence of the linear feature $p(\mathcal{Z})_i - p_0(\mathcal{Z})$ in (7), we have no guidance on the functional form we should adopt to the feature n_i. It is also likely that n_i has non-linear effects, e.g., a saturation level, when additional increments in an already large n_i does not increase $\mathbb{P}(Y_i=1)$. This is the justification for the adoption of a flexible non-parametric fit.

3.3 Model 2: Conditional Spatial Logistic Model with Matched Case-Control Individuals

Model (6) has one unappealing aspect, the presence of the nuisance offset $g(n_i)$. The total number of tweets affects the probability of ever posting a dengue-related tweet but we have no epidemiological interest in this relationship. Its presence is merely to control for the confounding variable n_i. However, we can get rid of this term if we match each dengue-labeled individual (individuals with $Y_i=1$, called cases) to other individuals with the same number of tweets and with $Y_i=0$ (called control individuals).

For each dengue-labeled case individual i with n_i tweets, we sample k non-diseased control individuals, all of them also with

n_i tweets. That is, we select k individuals among the subset of those with number of tweets equal to n_i and with their Y variable equal to zero. This matched sample will be represented by $\mathbf{S}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{ik})$ where $Y_{i0} = 1$ and $Y_{ij} = 0$ for $j > 0$. The individuals in the vector $\mathbf{S}_i$ have the same number n_i of tweets but differ on the locations of these tweets. Therefore, for any region Z, the feature $p(\mathcal{Z})_{ij}$ is likely to have different values for different values of j. The objective of this matching is to ensure comparability between cases and controls, reducing the systematic differences in the probability (6) due to the number of tweets.

Considering the matched sample $\mathbf{S}_i$, we know that only one of them has the *case* label and the others have the *control* label. We will consider the conditional probability that only the first element of $\mathbf{S}_i$ receives the *case* label given $Y_{i0} + Y_{i1} + \ldots + Y_{ik} = 1$. To avoid excessive notation, we drop the conditioning events $n_i, p(\mathcal{Z})_i$ from $\mathbb{P}(Y_i = 1 | n_i, p(\mathcal{Z})_i)$ writing it simply as $\mathbb{P}(Y_i = 1)$. We have

$$\mathbb{P}\left(Y_{i0} = 1 \middle| \sum_j Y_{ij} = 1\right) = \frac{\mathbb{P}(Y_{i0} = 1) \prod_{j=1}^{k} \mathbb{P}(Y_{ij} = 0)}{\sum_C \prod_{j=0}^{k} \mathbb{P}(Y_{ij} = a_j)} \tag{8}$$

where $C = \left\{(a_0, a_1, \ldots, a_k) : a_j \in \{0, 1\}, \sum_j a_j = 1\right\}$. Substituting (6) in (8), we find that:

$$\mathbb{P}\left(Y_{i0} = 1 \middle| \sum_j Y_{ij} = 1\right) = \frac{\exp(\beta\, p(\mathcal{Z})_{i0})}{\sum_{j=0}^{k} \exp(\beta\, p(\mathcal{Z})_{ij})} \tag{9}$$

This conditional probability is not affected by n_i any longer.

3.4 The Likelihood Function and Inference

We want to test the null hypothesis $H_0 : \lambda(x, y) = \lambda_{\text{all}}$ is constant versus the alternative hypothesis that there is region $\mathcal{Z}$ such that

$$H_1(\mathcal{Z}) : \begin{cases} \lambda(x, y) = \lambda_{\text{in}}, \ \forall (x, y) \in \mathcal{Z} \\ \lambda(x, y) = \lambda_{\text{out}}, \ \text{ otherwise} \end{cases}$$

with $\lambda_{\text{in}} > \lambda_{\text{out}}$. This alternative hypothesis is equivalent to having $\beta > 0$ in models (7) and (9).

The aim is to find the most likely region $\mathcal{Z}$ given the evidence provided by $\mathbf{S}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{ik})$ and the spatial locations of the tweets. For a given potential spatial cluster $\mathcal{Z}$, we can estimate the proportion of time that each individual spent inside $\mathcal{Z}$ (see section 3.5 for further details). Then for the unconditional spatial logistic model, the likelihood for the observed sample $Y_1, Y_2, \ldots, Y_n$ of binary variables is given by the logistic likelihood,

$$L^{m_1}(H_1, \mathcal{Z}, \beta) = \prod_i \mathbb{P}(Y_i = 1 | n_i, p(\mathcal{Z})_i)^{y_i} \mathbb{P}(Y_i = 0 | n_i, p(\mathcal{Z})_i)^{1-y_i}, \tag{10}$$

where the probability $\mathbb{P}(Y_i = 1 | n_i, p(\mathcal{Z})_i)$ is given in equation (7).

For the conditional model, for each disease case $Y_{i0} = 1$ with n_i tweets, we have a matched control sample of k individuals with $Y_{ij} = 0$ and the same number n_i of tweets. Considering a fixed region $\mathcal{Z}$, the conditional binomial likelihood for the observed samples $\mathbf{S}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{ik})$ for $i = 1, \ldots, N$ is given by

$$\begin{aligned} L^{m_2}(H_1, \mathcal{Z}, \beta) &= \prod_{i=1}^{N} \mathbb{P}(Y_{i0} = 1 | \sum_j Y_{ij} = 1, H_1, \mathcal{Z}) \\ &= \prod_{i=1}^{N} \frac{\exp(\beta p(\mathcal{Z})_{i0})}{\sum_{j=0}^{k} \exp(\beta p(\mathcal{Z})_{ij})} . \end{aligned} \tag{11}$$

For fixed $\mathcal{Z}$, the maximum likelihood estimator of β maximizes (10) or (11) and it is denoted by $\hat{\beta}(\mathcal{Z})$. The most likely zone $\hat{\mathcal{Z}}$ is finally given by

$$\hat{\mathcal{Z}} = \arg\max_{\mathcal{Z}} L^{m_k}(H_1, \mathcal{Z}, \hat{\beta}(\mathcal{Z})) \tag{12}$$

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 first model,

$$L^{m_1}(H_0) = \prod_i \mathbb{P}(Y_i = 1 \mid n_i)^{y_i} \mathbb{P}(Y_i = 0 \mid n_i)^{1-y_i}$$

where $\mathbb{P}(Y_i = 1 \mid n_i) = 1/(1 + \exp(g(n_i)))$ as $\beta = 0$ under H_0. Note that this likelihood function does not depend on $\mathcal{Z}$.

In the case of the second model, the likelihood for each matched sample is the probability of seeing $\mathbf{S}_i = (1, 0, \ldots, 0)$ when only one of the elements is selected with equal probability and therefore $L^{m_2}(H_0) = 1/(k + 1)^N$, not depending on $\mathcal{Z}$.

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

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

Next, we run a permutation test to obtain its associated p-value. In the case of the first model, we randomly permute the case and control labels (we 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 $T^{m_1,1}, T^{m_1,2}, \ldots, T^{m_1,nsim}$. The p-value is given by

$$p\text{-}value^{m_1} = \frac{\#\{T^{m_1,k} \geq T^{m_1,0},\ k = 0, 1, \ldots, nsim\}}{nsim + 1} , \tag{13}$$

which is approximately the proportion of permutation-based values $L(H_1, \hat{\mathcal{Z}}^{(j)}, \hat{\beta}(\hat{\mathcal{Z}}^{(j)}))$ that are larger than the observed value $L(H_1, \hat{\mathcal{Z}}^{(0)}, \hat{\beta}(\hat{\mathcal{Z}}^{(0)}))$.

In the second model, a restricted permutation distribution within each matched sample is carried out. Independently for each sample $(Y_{i0}, Y_{i1}, \ldots, Y_{ik})$, randomly assign the dengue label to one of them with equal probability. This random assignment generates a pseudo-dataset that is used as if they were the truly observed data. We recalculate the most likely zone $\hat{\mathcal{Z}}^{(1)}$ and its associated likelihood ratio $L(H_1, \hat{\mathcal{Z}}^{(1)}, \hat{\beta}(\hat{\mathcal{Z}}^{(1)}))$. Repeating this a large number *nsim* of times we obtain the sequence $\hat{\mathcal{Z}}^{(1)}, \ldots, \hat{\mathcal{Z}}^{(B)}$ and the empirical distribution of the test statistic $L(H_1, \hat{\mathcal{Z}}, \hat{\beta}(\hat{\mathcal{Z}}))$ under the null hypothesis. The p-value is again obtained using equation (13). The test is significant at the level $\alpha \in (0, 1)$ if p-value $< \alpha$. When either test is significant, the most likely zone is given by the corresponding maximizing argument $\hat{\mathcal{Z}}$. We also identify secondary clusters, that is, regions with p-values smaller than α that do not intersect with the most likely region zone $\hat{\mathcal{Z}}$. This non-intersecting restriction is necessary to avoid finding anomalous regions that are only slightly different from each other.

3.5 Implementation Issues

Estimating $p(\mathcal{Z})_i$: In order to avoid the instability due to small numbers when an individual has few tweets, $p(\mathcal{Z})_i$ is estimated from the number of tweets n_i in region $\mathcal{Z}$ using a smoothed Maximum Likelihood Estimator:

$$p(\mathcal{Z})_i = (\rho p_0(\mathcal{Z}) + (n_i \in \mathcal{Z}))/(\rho + n_i), \tag{14}$$

where $p_0(\mathcal{Z})$ is the mean of $p(\mathcal{Z})_i$ over the entire dataset. Thus,

$$p(\mathcal{Z})_i - p_0(\mathcal{Z}) = \frac{n_i \in \mathcal{Z} - n_i p_0(\mathcal{Z})}{\rho + n_i} = \frac{n_i}{\rho + n_i}\left(\frac{n_i(\mathcal{Z})}{n_i} - p_0(\mathcal{Z})\right).$$

Scanning Regions: Similarly to [20], we represent the spatial region under analysis as a rectangular grid of size $K_1 \times K_2$ cells. Then, we consider the set of rectangular regions on the grid as our search regions $\mathcal{Z}$. In other words, we evaluate a subset of grid cells if and only if the resulting region is rectangular and if its total area is smaller than a pre-defined percentage of the total grid area.

4 EXPERIMENTAL ANALYSIS

First, we apply our methods on semi-synthetic datasets. This way, we are aware of the actual ground truth information to evaluate the performance of our methods and to compare their performance with competing baselines. The datasets are semi-synthetic because we used actual mobility patterns from Twitter data and simulated only the individuals' labels Y_i (case or control) to obtain more realistic data, as described below.

4.1 Simulation Setup

We selected Campinas, a city in Southeast Brazil, to perform the experiment. The data were collected as described in section 5.1. At total, the Campinas dataset has 3,278 Twitter users issuing 456,761 messages between Jan. 1 and Dec. 31, 2015. We set a base grid of size 32×32 to scan the map. We selected an arbitrary region $\mathcal{Z}$ to be the simulated area of increased infection risk. Next, we generated individuals' labels as cases or controls according to the model given by Equation (7). We used the offset values $g(n_i)$ computed according to the label distribution in Section 5.3, for the same city and users. We vary the effect size as $\beta = \{0, 0.5, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4, 5\}$ and for each value we generated 50 simulations. The greater the value of β, the more users are labeled as cases. For $\beta = 0$, the mean number of users labeled as cases was 41.96 (1.28% of total users) and for $\beta = 5$ the mean was 109.48 (3.34% of total users).

4.2 Baseline Methods:

We compare our algorithms with four competing baselines: two variants of the standard Bernoulli spatial scan statistics [12] and two recently proposed models for cluster detection from trajectories [26]. Below, we briefly describe each baseline:

Bernoulli Scan 1 (BS1): We reduce the set of tweets from each individual to a single point by computing the single most common tweet location. Then, we consider the total number of case and control individuals for each location.

Bernoulli Scan 2 (BS2): We ignore the individuals issuing the tweets and just consider the total numbers of tweets from case and control individuals for each location.

Infection Model: It searches for the most likely region where a person gets infected when visiting.

Visit Model: It searches for the most likely region where case individuals go more frequently than control individuals.

For both BS1 and BS2, we apply the standard Bernoulli spatial scan to the reduced data as in [12]. Details of the Infection and Visit models are presented in [26].

4.3 Experimental Settings

We used the SaTScan software (available at http://satscan.org) along with the R package `rsatscan`, to run the Bernoulli scan baselines. Also, we used the same base grid and set the maximum cluster size to 20% of the population. The Infection and Visit models search over the same set of regions as our methods. Also, for the Infection and Visit models we sample controls as three times the number of cases as indicated in Souza et al. [26]. The number of Monte Carlo replicas were set to 999 and the significance level to $\alpha = 0.05$.

4.4 Simulation Results

4.4.1 *Measuring the spatial accuracy.*

We computed the average spatial accuracy, i.e., the degree of overlap between true (true $\mathcal{Z}$) and detected clusters (detected $\mathcal{Z}$) for each method as follows:

$$\text{spatial overlap} = \frac{\text{true } \mathcal{Z} \cap \text{detected } \mathcal{Z}}{\text{true } \mathcal{Z} \cup \text{detected } \mathcal{Z}}. \tag{15}$$

Spatial overlap is a measure of detection accuracy and measures how well a method can identify the exact true cluster. We want to assess whether our best guess at the region of elevated risk (i.e., our top-1 detected cluster) is better than that of all competing methods across simulations, as measured by overlap with the true cluster. Figure 2 shows the spatial overlap with the true cluster averaged over the 50 simulations for each method as a function of β. Although BS1 shows somewhat better spatial accuracy for small values of β, these signals are too subtle for any of the methods to perform well, as shown by our detection power comparison below (Fig. 3). As the effect size β increases to levels where the cluster is detectable, our methods outperform the competitors by a substantial margin.

4.4.2 *Comparing detection power against Bernoulli spatial scan variants.*

We measured the detection power at a significance level $\alpha = 0.05$. We count a region as "detected" if both: (i) the detected cluster's score is above the 95th percentile of the highest-scoring detected clusters for the simulated datasets generated under H_0, and (ii) if the detected cluster's overlap with the true cluster is at least x%, where the overlap is computed as in eqn. (15). Figure 3 shows the power curves for our models and both Bernoulli scan baselines when detecting the exact true cluster, i.e., overlap = 100% (left), and when detecting a cluster with overlap ≥ 60% (right). Notice

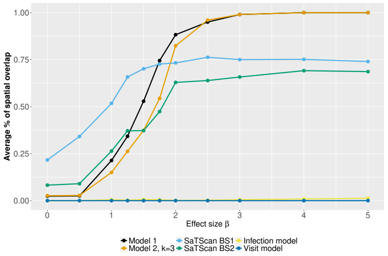

Figure 2: Average % of spatial overlap with true cluster over the 50 simulations for each method.

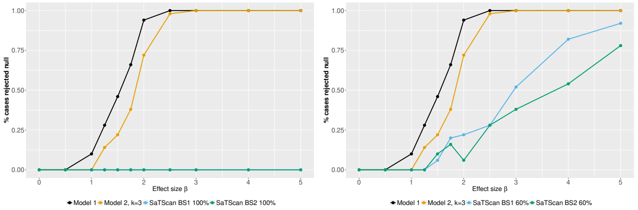

Figure 3: Power at $\alpha = 0.05$ for Model 1, Model 2, BS1 and BS2 for overlap $= 100\%$ (left) and $\geq 60\%$ (right).

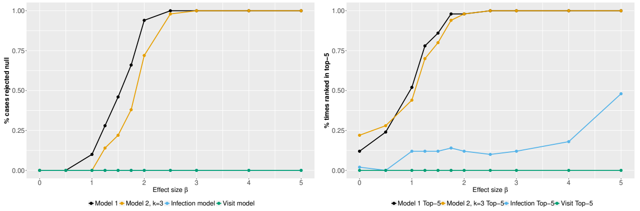

Figure 4: Left: power at $\alpha = 0.05$ for Model 1, Model 2, infection and visit models. Right: Percentage of times any region that overlaps the true cluster is ranked among top-5 scores.

that Model 1 and Model 2 dominate all the Bernoulli spatial scan variants, with Model 1 performing better, regardless of the overlap considered. Both BS1 and BS2 were unable to detect the exact true region for all values of β (power ≈ 0 for overlap $= 100\%$). When we consider spatial overlap $\geq 60\%$, the Bernoulli scan variants improved their results, but are still dominated by our methods.

4.4.3 Comparing detection power and ranking against competing models. We compare the detection power of our methods against the Infection and Visit models [25, 26]. In Figure 4 (left plot), we show the power curves for detecting the exact true cluster, i.e., overlap $= 100\%$. Both Infection and Visit models were not able to detect the exact true region for the considered values of β. Even considering overlap $\geq 60\%$, neither competing model found enough evidence to say any region was significant at $\alpha = 0.05$, while our models achieved consistently high performance. Thus we also considered a different metric to compare our models and the competing baselines. In Figure 4 (right plot), we considered the percentage of times the models ranked any region that overlaps with the true $\mathcal{Z}$ among the top-5 highest scores, regardless of whether it was significant at $\alpha = 0.05$. Considering this metric the Infection model has slightly improved performance, being able to rank a region that overlaps the true region among its top-5 scores, while the Visit model still performed poorly. Our two models had similar, high performance, with Model 2 beating Model 1 for smaller values of β.

4.4.4 Sensitivity analysis on the number of matched controls. We evaluate the sensitivity of Model 2 to the number of control individuals k matched to each case individual. We set $k = \{1, 2, 3, 4, 5\}$ and run Model 2 over the same datasets as above. Figure 5 depicts the power at $\alpha = 0.05$ to detect the true $\mathcal{Z}$ (left plot) and also the power to detect any $\mathcal{Z}$ (right plot), for each value of k. We note that, in both cases, even with $k = 1$ (i.e., a single control matched to each case), the algorithm was rapidly able to detect the true region. Detection power increases with increasing values of k, approximating the performance of Model 1. This result shows that Model 2 has good detection power and is robust to the choice of k, making it particularly useful when the offset term $g(n_i)$ is difficult to estimate.

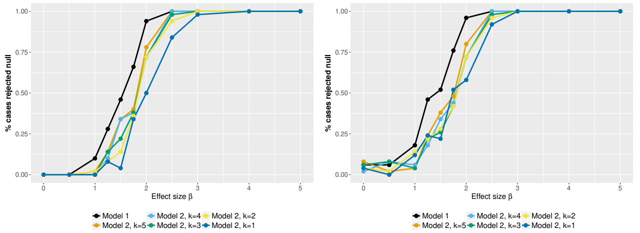

Figure 5: Sensitivity to parameter k: power at $\alpha = 0.05$ increases with increasing matching sample size.

5 CASE STUDIES

We consider two case studies using real-world data from Twitter. For the Airport application (§5.3) we know in advance the actual spatial clusters to be detected. For the Dengue application (§5.4), there is no ground-truth data and thus we perform a qualitative evaluation of the results.

5.1 Dataset Description

Our geolocated data were collected through the Twitter Streaming API[2]. The collection period was from January 1 to December 31, 2015, during which we were able to crawl a total of 106,784,441 Twitter messages geo-tagged with lat/long GPS coordinates. To do so, we set a geographic boundary box covering the Brazilian territory, filtered out the messages issued from outside Brazil, and assigned all remaining tweets to their corresponding valid municipality, therefore generating one dataset per city.

5.2 Experimental Setup

In order to apply Model 1, we need to estimate the offset $g(n_i)$ given by eqn. (5). To do so, we split the values of n_i in ranges and compute the odds of being a case in each range. The ranges were

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

set to powers of 2 from $n_i = 16$ up to $n_i = 2048$. We filtered out individuals with $n_i < 16$ as we would be insufficiently certain about their typical locations and their "infection" status. Also, we filtered out users with $n_i > 2048$ to potentially avoid bots. After computing the odds of being a case in each range, we fit a function to learn the relationship between n_i and $g(n_i)$. In the Airport application (§5.3) we used locally weighted linear regression (loess) for this task. In the Dengue application (§5.4) we used a linear model as the log-log graph suggests a power law relationship between n_i and $g(n_i)$. For Model 2, we set the number of matched control individuals to $k = 3$. We used the same matching for the Infection and Visit Models. For all models we used a base grid of 32 × 32 in the Airport application and 40 × 40 in the Dengue application. Also, we set the number of Monte Carlo replicas to $B = 499$.

5.3 Airport Application

In this first application, we define the case individuals as those Twitter users mentioning the keyword *airport*. Our reasoning behind this keyword choice is that, we can anticipate some of the potential "infection" clusters: airport areas, tourist places where people that are visiting the city (maybe arriving through the airport) are likely to go, and places where people can take transportation to the airports, to name a few. The control group is composed by all other individuals who never mentioned the keyword. Our goal is to contrast the mobility patterns of case and control individuals to search for relevant spatial clusters. Notice that individuals who used the keyword but did not go to the airport or associated locations are also labeled as cases, and conversely, individuals who went to the airport and associated locations but never used the keyword are labeled as controls. Moreover, the tweet mentioning "airport" may or may not have occurred at the airport.

We selected the city of Campinas to perform our analysis. First, we search the Campinas dataset to uncover the case individuals (users mentioning the keyword *airport*) and control individuals. Then, we retrieve all messages issued by each individual, regardless of whether or not these messages contain the keyword, to build their mobility patterns. Out of the 3,278 distinct individuals tweeting from Campinas in our dataset, 223 of them were labeled as cases (with 42,579 tweets) and the remaining 3,055 as controls (with 414,182 tweets).

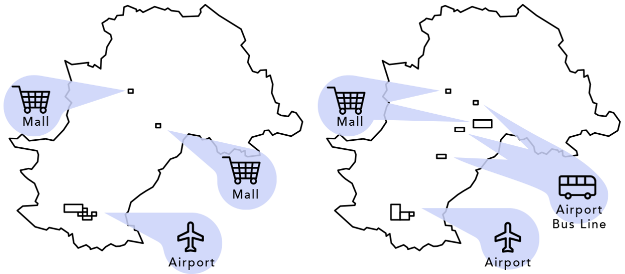

Figure 6: Regions detected in Campinas by Model 1 (left) and Model 2 (right) and corresponding explanations.

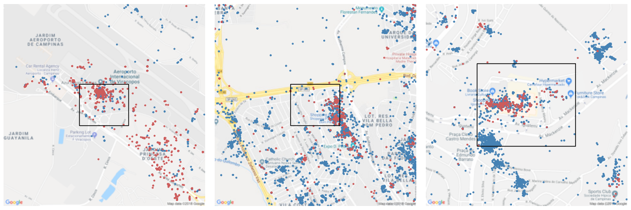

Figure 7: Zoom in to top-3 regions detected by Model 1 (black rectangles) in Campinas for the airport application.

We followed the experimental setup described in §5.2 to run the algorithms. Model 1, Model 2 and the Infection model were able to detect significant regions. The Visit model did not find significant regions. A thorough verification showed that all regions detected by our models are consistent with our expected clusters. Model 1 detected 8 significant regions: 6 in the Campinas international airport area and 2 over the area of two malls, containing a popular tourist attraction. Model 2 also detected 8 regions: 3 in the airport area, 2 over the same two malls as Model 1, and 3 regions in the exact path of the airport bus line. Figure 6 shows two maps of Campinas, each one with the regions detected by our models along with their descriptions mentioned above. The Infection model also detected one region located in the airport area (coincident with one of the regions detected by Model 2), but was not able to detect any other associated region. Table 1 shows numerical details on the top-3 regions detected by Model 1 and Model 2, as well as the single region detected by the Infection model. We also show the top-3 regions detected by Model 1 in Figure 7, with the positions of the tweets issued by case (red dots) and control (blue dots) individuals, and the area of the detected regions (black rectangles). On all three maps we can see a clear cluster of case individuals.

Table 1: Airport application results: LLR score, model parameter β, p-value, and numbers of case/control individuals and case/control tweets, for each detected area of elevated risk.

	LLR	β	p-value	#cas	#ctl	#tw_cas	#tw_ctl
Model 1	150.68	44.186	0.005	177	42	1,164	212
	23.39	134.821	0.005	14	4	221	32
	18.10	9.267	0.005	105	454	681	1,805
Model 2	155.58	123.721	0.005	182	53	1,326	393*
	14.71	96.009	0.005	17	8	268	106
	13.10	10.577	0.005	105	454	681	1,805

* Also found by Infection model (LLR = 1340.53, p-value = 0.01).

5.4 Dengue Application

As discussed in §1 above, our work is motivated by the search for spatial clusters associated with high dengue infection risk. These places are very hard to identify since the required information to do so is scarcely available or infeasible to obtain. For example, we would wish to (but do not) know data such as the infection rate in each area, infected mosquito population, and biting rates at each potential location. That is where we expect our methodology and algorithms to be most useful, revealing potential high-risk areas which could then be validated by public health.

We label each individual as case or control based on the content of their tweets. In order to find individuals presenting a dengue infection episode, we follow [25] and search for all tweets presenting the keywords *dengue* or *Aedes*. Differently from the previous application, we are interested in retaining only messages with strong evidence of dengue infection and not all the individuals mentioning the keyword. Therefore, we used a set of tweets manually labeled into five categories (personal experience, information, campaign, opinion and irony/sarcasm, following the taxonomy of [4]) to train a classifier and automatically labeled the remaining tweets. We use a Lazy Associative Classifier [30] and refer the reader to [25] for labeling and classification details. The group of case individuals is defined as those users who had at least one tweet in the *personal experience* category. Previous works [8, 27] have shown a high correlation between Twitter mentions of personal experience with dengue and official reports of dengue incidence, therefore we retain only these messages. The control individuals group is composed by the remaining users. Similarly to the previous application, each individual's mobility patterns are composed of locations from all tweets issued by that user. We note that tweets occurring after the dengue infection may still provide useful information about an individual's typical mobility patterns, since they may have been infected at that location at another time.

We followed the experimental setup described in §5.2 in order to run the algorithms. We selected Sorocaba, a city in Southeast Brazil with ∼650,000 inhabitants to perform the analysis. The city was selected because it was highly affected by the 2015 dengue outbreak in Brazil, reporting <400 dengue cases in 2014 and >55,000 cases in 2015.[3]. The Sorocaba city dataset has 74 case individuals with 30,232 tweets and 1,521 control individuals with 169,399 tweets. Model 1 detected 9 significant regions at $\alpha = 0.05$, and the Infection model detected 2 significant clusters. Model 2 identified one cluster that was significant at $\alpha = 0.1$ but not $\alpha = 0.05$; such regions may also be worth monitoring depending on available public health resources. The Visit model was not able to detect any regions in this dataset. Table 2 shows numerical details for the regions found by all three models (for Model 1 we show only the top-2 regions).

The most difficult part of searching for places with increased risk of infection by dengue (and other vector-borne diseases) is that dengue is just one of many infectious diseases with a well-known etiology but a huge number of uncertain and difficult to obtain parameters that quantify factors such as infected mosquito population, likelihood of being bitten by an infected mosquito, and human movement in the mosquito-infested areas. Therefore, we take a closer look at the regions detected by Model 1 and shown in Figure 8. A detailed inspection of each region revealed that only a small portion of the detected regions were located in areas that are mainly residential; several non-residential places were detected, such as a hospital, college campus, airport, and parks. Standard surveillance systems using only the residence address would not have been able to detect such regions, showing the potential for our approach to add to public health understanding of dengue risk.

Table 2: Dengue application results: LLR score, model parameter (β for Models 1 and 2, $r/\bar{r}$ for Infection model), p-value, and numbers of case/control individuals and case/control tweets. Top, middle, and bottom partitions show detected areas of elevated risk for Model 1, Model 2, and Infection model respectively.

LLR	param	p-value	#cas	#ctl	#tw_cas	#tw_ctl
17.825	0.184	0.005	22	30	2,190	122
17.755	2.719	0.005	9	4	200	5
6.216	11.785	0.09	10	4	182	325
446.804	0.04 / 0.01	0.002	2	3	11	10
446.946	0.04 / 0.01	0.002	3	150	8	16

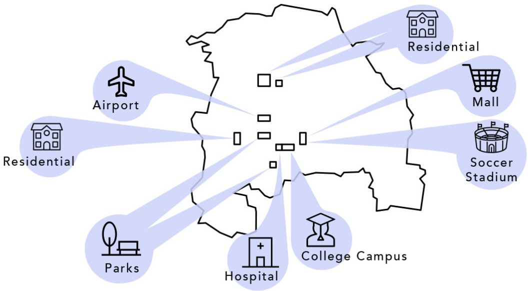

Figure 8: Regions detected in Sorocaba by Model 1 and their corresponding explanations.

6 REPRODUCIBILITY

We have built (and will make freely available upon request) a custom R package to run our algorithms. Our code takes as inputs the features used by each model, computed for all individuals under analysis over all regions to be scanned. The outputs consist of a set of locations, the corresponding score, estimated parameters and associated p-value. The provided package also contains all the data used in this paper, including both simulated and real-world datasets, as well as the full dataset described in §5.1, thus ensuring full reproducibility of our results.

7 CONCLUDING REMARKS

Identifying places where people have higher risk of being infected, rather than focusing on residential address locations, may be key to surveillance, especially for infectious diseases where human mobility plays a significant role (e.g., dengue infection). Being able to pinpoint such regions allows public health officials to focus prevention and mitigation actions, such as mosquito control, where they are most needed.

We proposed two new spatial scan methods (the unconditional and conditional spatial logistic models) to search for spatial clusters

[3] http://portalarquivos2.saude.gov.br/images/pdf/2016/janeiro/07/2015-svs-be-pncd-se48.pdf (in Portuguese)

of increased infection risk in mobility patterns. As our experiments demonstrate, the stochasticity of mobility data causes typical spatial cluster detection tools, such as the traditional spatial scan statistic, to fail. Moreover, 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. Our methods add to the set of tools that both public health researchers and practitioners have available to search for spatially localized infection risk clusters using readily available Twitter data. We expect that our methods will also be useful to other public health surveillance problems where individuals' movement data can bring relevant information.

ACKNOWLEDGMENTS

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

[1] R. Assunção, M. Costa, A. Tavares, and S. Ferreira. 2006. Fast detection of arbitrarily shaped disease clusters. Stat. Med. 25, 5 (2006), 723–742.

[2] S. Bhatt et al. 2013. The global distribution and burden of dengue. Nature (2013).

[3] F. Chen and D. B. Neill. 2014. Non-parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs. In SIGKDD. 1166–1175.

[4] C. Chew and G. Eysenback. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. Plos One 5 (2010).

[5] M. Costa, R. Assunção, and M. Kulldorff. 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. CSDA 56, 6 (2012).

[6] E. Cromwell et al. 2017. The relationship between entomological indicators of Aedes aegypti abundance and dengue virus infection. PLoS NTDS 11, 3 (2017).

[7] L. Duczmal and R. Assunção. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. CSDA 45, 2 (2004), 269–286.

[8] J. Gomide, A. Veloso, W. Meira Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In Proc. of the ACM WebSci Conference.

[9] W. Herlands, E. McFowland, A. Wilson, and D. Neill. 2018. Gaussian Process Subset Scanning for Anomalous Pattern Detection in Non-iid Data. In AISTATS.

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

[11] G. M. Jacquez, J. Meliker, G. AvRuskin, P. Goovaerts, A. Kaufmann, M. Wilson, and J. Nriagu. 2006. Case-control geographic clustering for residential histories accounting for risk factors and covariates. IJHG 5, 1 (2006), 32.

[12] Martin Kulldorff. 1997. A Spatial Scan Statistic. Comm. in Stat. - Theory and Meth. 26, 6 (1997), 1481–1496.

[13] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. 2006. An elliptic spatial scan statistic. Stat. Med. 25, 22 (2006), 3929–3943.

[14] A. Lawson et al. 1999. Disease mapping and risk assessment for public health. Wiley New York.

[15] M. Matheny, R. Singh, L. Zhang, K. Wang, and J. M. Phillips. 2016. Scalable spatial scan statistics through sampling. In 24th ACM SIGSPATIAL.

[16] E. McFowland, S. Speakman, and D. Neill. 2013. Fast generalized subset scan for anomalous pattern detection. JMLR 14, 1 (2013), 1533–1561.

[17] N. Murray, M. B Quam, and A. Wilder-Smith. 2013. Epidemiology of dengue: past, present and future prospects. Clinical Epidemiology 5 (2013), 299–309.

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

[19] Daniel B Neill and Wilpen L Gorr. 2007. Detecting and preventing emerging epidemics of crime. Advances in Disease Surveillance 4, 13 (2007).

[20] Daniel B. Neill and Andrew W. Moore. 2004. Rapid Detection of Significant Spatial Clusters. In Proc. of the 10th ACM SIGKDD. 256–265.

[21] World Health Organization et al. 2004. The World health report: 2004: changing history. Geneva: World Health Organization.

[22] World Health Organization, Special Programme for Research, and Training in Tropical Diseases. 2009. Dengue: guidelines for diagnosis, treatment, prevention and control. World Health Organization.

[23] L. Shi and V. Janeja. 2009. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). In SIGKDD. 767–776.

[24] S. Somanchi, D. B. Neill, and A. Parwani. 2018. Discovering anomalous patterns in large digital pathology images. Statistics in medicine 37, 25 (2018), 3599–3615.

[25] R. CSNP Souza, R. Assunção, D. Oliveira, D. B. Neill, and W. Meira Jr. 2018. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. Spatial and Spatio-temporal Epidemiology (2018).

[26] R. CSNP. Souza, R. Assunção, D. M. Oliveira, D. E. F. Brito, and W. Meira Jr. 2016. Infection Hot Spot Mining from Social Media Trajectories. In ECML/PKDD.

[27] R. CSNP Souza, D. de Brito, R. Cardoso, D. de Oliveira, W. Meira Jr., and G. Pappa. 2014. An Evolutionary Methodology for Handling Data Scarcity and Noise in Monitoring Real Events from Social Media Data. In IBERAMIA. 295–306.

[28] Steven Stoddard et al. 2013. House-to-house human movement drives dengue virus transmission. PNAS 110, 3 (2013), 994–999.

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

[30] A. Veloso, Wagner Meira Jr., and M. J. Zaki. 2006. Lazy Associative Classification. In Proc. of the International Conference on Data Mining. 645–654.

[31] M. Wu, X. Song, C. Jermaine, S. Ranka, and J. Gums. 2009. A LRT framework for fast spatial anomaly detection.. In KDD. 887–896.