Spatial and Spatio-temporal Epidemiology 29 (2019) 163–175

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Original Research

Where did I get dengue? Detecting spatial clusters of infection risk with social network data

Roberto C.S.N.P. Souza[a,*], Renato M. Assunção[a], Derick M. Oliveira[a], Daniel B. Neill[b], Wagner Meira Jr.[a]

[a] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[b] Center for Urban Science and Progress, New York University, New York, NY, United States

ARTICLE INFO

Article history:
Received 7 December 2017
Revised 13 June 2018
Accepted 14 November 2018
Available online 1 December 2018

Keywords:
Spatial cluster detection
Disease surveillance
Dengue
Social media data
Mobility data
Scan statistics

ABSTRACT

Typical spatial disease surveillance systems associate a single address to each disease case reported, usually the residence address. Social network data offers a unique opportunity to obtain information on the spatial movements of individuals as well as their disease status as cases or controls. This provides information to identify visit locations with high risk of infection, even in regions where no one lives such as parks and entertainment zones. We develop two probability models to characterize the high-risk regions. We use a large Twitter dataset from Brazilian users to search for spatial clusters through analysis of the tweets' locations and textual content. We apply our models to both real-world and simulated data, demonstrating the advantage of our models as compared to the usual spatial scan statistic for this type of data.

1. Introduction

Public health surveillance agencies traditionally use data sources including Emergency Department chief complaints, electronic medical records, microbiological data, and over-the-counter medication sales to identify emerging spatial clusters of disease. The delay and cost of obtaining these data has prompted many initiatives to take advantage of the timely, ready and cheap availability of health information on the Web. Personal health conditions are a common conversational topic in online social networks. Such data can provide a continuous and useful source of information for health agencies to perform real-time surveillance (Paul and Dredze, 2011; Prieto et al., 2014). The guiding principle in this case is to consider people as sensors and their corresponding messages as an indicator of the occurrence or intensity of the monitoring aim, such as disease outbreaks (Chen and Neill, 2014).

Health surveillance systems usually identify high risk places based only on the residence address or the working place of diseased individuals. This approach ignores a multitude of exposures the individuals are daily subject to and therefore provides little information about the actual places where people are infected, the truly important information for disease control. The increasing availability of geolocated data in online platforms offers a unique opportunity: in addition to identifying diseased individuals, we can also follow them in time and space as they move on the map. Incorporating the mobility of individuals into spatial analysis requires the development of new models that can cope with this type of data in a principled way and efficient algorithms to deal with the ever growing amount of data.

In this work, we give a contribution in this direction. We exploit geolocated data from online social networks

* Corresponding author.
E-mail addresses: nalon@dcc.ufmg.br (R.C.S.N.P. Souza), assuncao@dcc.ufmg.br (R.M. Assunção), derickmath@dcc.ufmg.br (D.M. Oliveira), daniel.neill@nyu.edu (D.B. Neill), meira@dcc.ufmg.br (W. Meira Jr.).

https://doi.org/10.1016/j.sste.2018.11.005
1877-5845/

to detect geographic clusters of dengue infection. Dengue is an infectious disease that is currently a major concern for public health officials, particularly in developing countries (Bhatt, 2013). We crawled a large collection of GPS-annotated data from Twitter. Individuals presenting a personal experience with the disease ("cases") are identified based on the sentiment conveyed in the content of their messages. We follow them in time and space to build their spatial trajectories, i.e., we retrieve a sequence of spatial locations that provide an estimate of individuals' movements on the map. We also build the trajectories of a baseline population ("controls"). Our goal is to contrast observed mobility patterns for case and control individuals in order to detect localized regions with higher risk of being infected by dengue. Identifying places where people have higher risk of being infected by the disease may be key to surveillance, particularly for vector-borne diseases such as malaria and dengue, allowing public health officials to focus mitigation actions.

The main contributions of this paper are as follows:

- We present two probabilistic models to search for spatial regions of higher risk of infection by dengue disease using movement data from social media.
- We thoroughly describe a methodology designed to identify case and control individuals and to extract their trajectories from social media data.
- We present results of applying our models to geolocated Twitter data considering two large Brazilian cities, showing the effectiveness of our methods.
- We compare our models to the Bernoulli spatial scan statistic on simulated data to demonstrate that assigning individuals to a single spatial position may provide misleading conclusions in some situations.

The remainder of this paper is organized as follows. In Section 2, we revisit the traditional spatial cluster detection problem and discuss the most common methods to solve the problem as well as some extensions. In Section 3 we give a background on dengue disease and discuss how people's movement can be an important factor when searching for spatial clusters of dengue infection. Section 4 presents two models to detect high risk regions based on trajectories of case and control populations. In Section 5, we thoroughly describe each step of our methodology to obtain spatial trajectories of individuals from online social networks. In Section 6, we apply the presented methods to Twitter data in order to search for dengue infection clusters. We also discuss the results obtained by both models and present a comparison with the Bernoulli spatial scan statistics. Finally, Section 7 provides some discussion and concluding remarks.

2. Spatial cluster detection

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 a 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, 1997; Kulldorff and Nagarwalla, 1995) 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:

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

where Data is a generic name for the observed data that is specified according to the model, e.g., the distribution of disease cases over space and time.

For the Bernoulli spatial scan the alternative hypothesis $H_1(Z)$ assumes that the probability of being a case within $Z \in \mathcal{Z}$ is higher than outside Z, and the null hypothesis H_0 assumes complete spatial randomness, i.e, each individual is equally likely to be a case everywhere in the map. After maximizing Eq. (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 derivation of the likelihood ratio test for the Bernoulli model can be found in Kulldorff (1997).

A major application of spatial scan statistics and many of its extensions is the detection of disease clusters to suggest risk factors, to focus preventive efforts, and for outbreak monitoring (Brooker et al., 2004; Hjalmars et al., 1996; Jones et al., 2012; Mostashari et al., 2003). However, they have also been applied to several other tasks, such as the identification of hot spots zones based on the geographical locations of crime events (Nakaya and Yano, 2010; Neill and Gorr, 2007) or traffic accidents (Shi and Janeja, 2009). Also, in order to overcome the limitation of a rigid (circular) scanning window, the set of spatial zones was enlarged by allowing elongated (Neill and Moore, 2004), elliptical (Kulldorff et al., 2006), and irregularly-shaped regions (Assunção et al., 2006; Costa et al., 2012; Duczmal and Assunção, 2004; Tango and Takahashi, 2005).

In all this large body of work, there has been one invariant aspect of the spatial characterization: there is one and only one spatial position associated with each individual data item, whether that location represents a pixel in a medical image, as in Somanchi et al. (2018), or a random spatial event, such as a crime or accident location. In spatial epidemiology, searching for environmental putative sources of infection or disease, in a few cases there have been two positions associated with each individual, their residential and working place addresses. Such an approach ignores a multitude of exposures that individuals may be subject to during their daily routine. Indeed, as we will discuss in Section 3, there have been several studies pointing out that most people get infected away from home (Stoddard et al., 2013; 2009).

3. Dengue overview

Dengue is an emerging mosquito-borne viral disease. Estimates on the number of global infections per year have ranged from 50 to 100 million cases counting clinically manifested infections up to almost 400 million cases including asymptomatic carriers (Bhatt, 2013; Murray et al., 2013). Such a large range on the estimated incidence of the disease clearly indicates that the true numbers are difficult to assess, mostly due to misdiagnosis and underreporting. The World Health Organization (WHO) estimates that almost half of the world's population is at risk of infection with dengue viruses, the majority being concentrated in the South and Central Americas, Asia and Pacific regions.

With four known serotypes, dengue may vary from severe flu-like illness to a potentially lethal complication known as hemorrhagic dengue. The global incidence of the disease keeps growing both in number and severity of cases, presenting approximately 20,000 associated deaths occurring annually (Murray et al., 2013). Since there is no currently approved, effective and broadly available vaccine to protect the population against the virus, epidemiological surveillance and effective vector control are still the mainstay of dengue fever prevention.

The main vector for transmission of dengue virus is the *Aedes aegypti* mosquito. Several entomological indicators have been proposed to quantify the abundance of *Aedes aegypti* since its monitoring was first employed for yellow fever control (Cromwell et al., 2017). The rationale behind such indicators is that the greater the mosquito population, the higher the risk of dengue transmission and therefore intervening to reduce the vector abundance consequently decreases the number of infections. However, recent studies have shown that there is no accurate correlation between vector prevalence and dengue transmission (Bowman et al., 2014; Cromwell et al., 2017). In addition, mosquito prevalence data is costly to obtain, particularly at large scale.

In fact, dengue has a huge amount of uncertain and difficult to obtain parameters driving the disease. Human mobility is one of the key factors, especially due to the mosquito day-biting habit (Stoddard et al., 2013; 2009). In this sense, attaching each individual to a single location, their home address, may be a poor indicator of the regions with higher level of interaction between humans and infected vectors. Being able to identify the most risky places would greatly benefit infectious disease surveillance by targeting preventive efforts and mitigation actions where they are most needed.

3.1. Dengue in Brazil

Brazil reports more cases of dengue than any other country.[1] In 2015 the Brazilian Ministry of Health reported approximately 1.6 million cases of dengue infection. This number represents a rate of 788 cases per 100 thousand inhabitants, well above the red line indicated by the WHO, which is 300 cases. In addition, 839 deaths were confirmed to be caused by dengue in the same period.[2] The Brazilian disease surveillance system is almost entirely manual and relies on the ability to observe early cases of dengue for each location and time period. This process usually results in long delays for data acquisition. Despite the huge amount of resources spent for surveillance and prevention actions, dengue still challenges Brazilian health services and policy makers. Previous studies leveraged online social network data to predict the incidence of dengue in Brazil (Gomide et al., 2011; Souza et al., 2014). However, they are not able to pinpoint high risk regions. We believe that our approach can bring significant contribution to the spatial epidemiology and surveillance of dengue.

4. Detecting spatial clusters from trajectories

We use Fig. 1 to explain the problem. In the left-hand side, each individual is indexed by a number i and has a set of n_i spatial positions. In this case, the positions are given by geolocated tweets. The tweets from a single person are connected by line segments. The individuals are additionally labeled by two colors according to their status: dengue case (in red) or control (in blue). The cases are those individuals who mentioned a personal experience with dengue in at least one tweet, as we will discuss in Section 5. The tweets which have mentioned personal experience with dengue are marked with a hatched shadow in Fig. 1. The figure also shows a spatial zone Z where the risk of becoming a case 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. If a candidate zone Z is easy to identify in this toy example, the difficulty with real data is much higher as the right hand-side of Fig. 1 demonstrates. We show a sample of tweets issued from the central area of Belo Horizonte, a city in the Southeast region of Brazil. With this realistic amount of data, it is obvious that simple visual inspection of the map is not effective and that a computer-based algorithm is necessary to find the most plausible spatial clusters.

The multiple locations associated with each individual, rather than the usual single location (such as their place of residence), leads us to consider two different models, which we term the *visit model* and the *infection model* (Souza et al., 2016). They are defined in terms of two events representing the individual becoming a case and tweeting from a certain zone Z, respectively. The models consider two different conditional probabilities: for a given individual i, while the visit model examines

$$p(Z) = \mathbb{P}(\text{individual } i \text{ tweets from } Z \mid \text{individual } i \text{ is a case}), \tag{2}$$

the infection model evaluates

$$r(Z) = \mathbb{P}(\text{individual } i \text{ is a case} \mid \text{individual } i \text{ tweets from } Z). \tag{3}$$

[1] http://www.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html.

[2] http://portalarquivos.saude.gov.br.

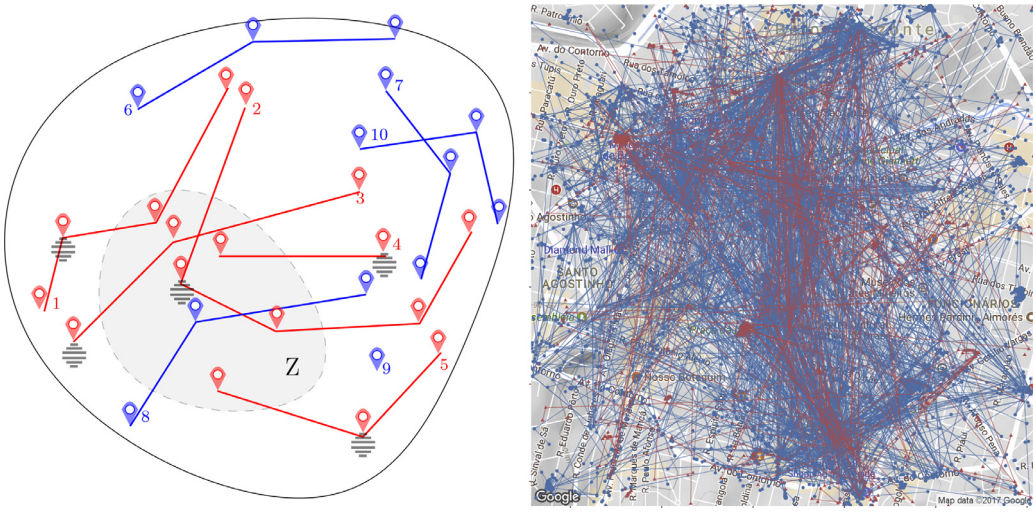


Fig. 1. *Left*: Schematic drawing of the problem showing a potential infection spatial cluster and trajectories of case (red) and control (blue) individuals. *Right*: trajectories of case and control individuals built based on a sample of tweets issued from the central area of Belo Horizonte in 2015. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, the models consider different aspects of the same problem, depending on which event we condition on.

To complete the specification of the models, let $\tilde{p}(Z)$ be the analogue of (2) for a control individual, and $r(\tilde{Z})$ the probability given by Eq. (3) but evaluated in $\tilde{Z}$, the region outside Z. Our interest resides only on zones where there is enough evidence to conclude that $p(Z) > \tilde{p}(Z)$ or $r(Z) > r(\tilde{Z})$.

Among the n_i tweets from the ith individual, let $V_{i,z}$ be the number inside the spatial cluster Z. The *visit model* considers the binary variables $\mathbb{1}[V_{i,z} \geq 1]$, i.e., the likelihood that individual i visits zone Z at any point during the study period. Its likelihood $\mathbb{P}(\text{Data} \mid \text{H}_1(\text{Z}))$ under the alternative hypothesis $H_1(Z)$ is given by the product of Bernoulli random variables defined for each individual. Let $\mathbb{1}[V_{i,z} \geq 1]$ indicate the event that the ith individual visits Z at least once. For a case individual, we have $V_{i,z} = 0$ if individual i never visits Z in his n_i tweets, which happens with probability $(1-p)^{n_i}$, and $V_{i,z} \geq 1$ with probability $1-(1-p)^{n_i}$. For a control individual, we have similar formulas with $\tilde{p}$ replacing p. Then the likelihood of the data is given by the product over all individuals, both cases and controls:

$$L_1(Z, p, \tilde{p}) = (1-p)^{\sum_{i=1}^{N} n_i \mathbb{1}[V_{i,z}=0]} (1-\tilde{p})^{\sum_{i=N+1}^{N+M} n_i \mathbb{1}[V_{i,z}=0]} \prod_{i=1}^{N} \left[(1-(1-p)^{n_i})^{\mathbb{1}[V_{i,z}\geq 1]} \right] \prod_{i=N+1}^{N+M} \left[(1-(1-\tilde{p})^{n_i})^{\mathbb{1}[V_{i,z}\geq 1]} \right], \quad (4)$$

where the first N individuals are cases and the last M are the control individuals. To simplify the expression, we dropped the zone Z from $p(Z)$ and $\tilde{p}(Z)$ writing simply p and $\tilde{p}$. The null model $\mathbb{P}(\text{Data} \mid \text{H}_0)$ in the denominator of Eq. (1) is obtained by making $p = \tilde{p}$ for all Z.

For the *infection model*, let the binary indicator $I_i = 1$ if the ith individual is a case, and let k_i be the individual's number of tweets in zone Z. Then we define

$$\pi(k_i, r, \tilde{r}) = \mathbb{P}(I_i = 1 | V_{i,z} = k_i) = 1 - (1-r)^{k_i}(1-\tilde{r})^{n_i - k_i},$$

and the likelihood under $H_1(Z)$ is given by:

$$L_2(Z, r, \tilde{r}) = \prod_{i=1}^{N+M} (\pi(k_i, r, \tilde{r}))^{I_i} (1 - \pi(k_i, r, \tilde{r}))^{1-I_i}. \quad (5)$$

For the null model H_0 in (5), similarly to the visit model, we take $r = \tilde{r}$ for all Z.

The most likely spatial cluster Z is found by first maximizing (4) over p and $\tilde{p}$ for the visit model and maximizing (5) over r and $\tilde{r}$ for the infection model for each fixed zone Z. Next, we maximize over Z to identify the highest-scoring (most significant) spatial clusters. The p-value of each cluster is then obtained by randomly permuting the cases and control labels among the individuals and recalculating the maximum likelihood ratio, as given by Eq. (1). After a large number of independent permutations, we have the empirical distribution of the maximum likelihood ratio under the null hypothesis, and the p-value can be obtained as the proportion of times the simulated values of the maximum likelihood ratio were larger than the observed value.

5. Dataset

In this section we thoroughly describe each step of our methodology designed to identify the case and control individuals and extract their trajectories from GPS-annotated social media data. The first step is the collection of geolocated Twitter data (Section 5.1). Next, we need to assign each message to a valid location based on its embedded spatial coordinates (Section 5.2). After that, we define the group of case individuals by filtering and analyzing the content of the tweets (Section 5.3). The individuals not selected in the previous task compose the control group. Finally, for individuals in each group, we build their trajectories by retrieving all geolocated messages they issued during the period of analysis (Section 5.4).

5.1. Data acquisition

The data used in our experimental analysis were acquired through the Twitter Streaming Application Programming Interface (API).[3] Twitter users are allowed to disclose their location in a number of different ways. They can fill in a free text field in their profile or Twitter can obtain and provide an approximate location based on the IP addresses. Tweets can also be geotagged with latitude-longitude GPS coordinates when tweeting from mobile devices with that feature enabled. While the first options are typically too coarse for a detailed spatial analysis (usually reporting the city or state where the user lives), the geotagged tweets allow us to track users' movement patterns with a reasonably good resolution. Therefore, in this study, we focus on geotagged tweets.

The Twitter API allows us to specify a geographic bounding box and collect the public tweets issued within that location together with their associated lat/long coordinates. The API also limits the crawling to a maximum of 1% of the total Twitter fire hose. However, this amount is just about the total volume of GPS-annotated posts (Sloan and Morgan, 2015), enabling us to collect the vast majority of the geotagged tweets within the bounding box.

We set a bounding box covering the Brazilian territory, defined by the points $[-33.751, -73.986]$ SW, $[5.265, -34.288]$ NE. Data was collected from January 1st, 2015 to December 31th, 2015. During this time period we were able to collect a total of 106,784,441 Twitter messages. All collected tweets are geotagged with lat/long GPS coordinates.

5.2. Location assignment

The geographic bounding box set to collect the data also includes regions outside Brazil. We filtered out the messages issued from these areas. Next, we need to assign each message coming from the Brazilian territory to a valid municipality. In Brazil, the decision process regarding dengue surveillance actions is under the responsibility of each town hall. Thus, performing our analysis for each Brazilian city separately can provide the responsible health officials with a list of potential high-risk areas inside their corresponding town.

We selected two municipalities to analyze based on the total number of dengue cases reported by the Brazilian Ministry of Health. Among the cities with more than 1 million inhabitants, we selected the two cities with the highest 2015 incidence rate of cases per 100 thousand inhabitants. Table 1 provides general information about the selected cities.

In order to process the location of each collected tweet we used the OpenStreetMap API[4] and retrieved the spatial polygons of all Brazilian cities. Then, for each tweet, we assign its lat-long information to the corresponding city by checking in which polygon it falls within. Table 1 presents the total number of collected tweets that were issued from within each of the selected cities.

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

[4] https://www.openstreetmap.org/about.

Table 1
Selected cities: #tweets is the total number of Twitter posts issued by users within the city; #reports is the total number of dengue cases in the city according to official reports; *Population* shows the number of inhabitants; *Rate* is the number of dengue cases per one hundred thousand inhabitants.

City	#tweets	#reports	Population	Incidence rate
Campinas	574,226	66,577	1,164,098	5719.2
Goiânia	566,114	74,097	1,448,639	5114.9

Table 2
Sentiment categories, the associated semantics and examples of real tweets (translated from Portuguese) belonging to each class.

Sentiment	Semantics	Tweets
Personal experience	Express dengue cases	"I am staying in bed. Got Dengue!"
Information	Carries some type of information	"Confirmed first case of dengue type 4."
Opinion	Express public opinion	"I hate this dengue mosquito-repellent smoke"
Campaign	Reinforces public campaigns	"Everyone against dengue! That's our fight!"
Irony/sarcasm	Jokes, sarcasm, or irony	"My social media is so quiet that it looks like breeding dengue water."

5.3. Textual content filtering and analysis

The content of geotagged tweets comprises a multitude of subjects. In order to find the individual cases, we need to check the content of the messages looking for evidences about a dengue infection for that particular user. This is not a straightforward task. For instance, despite having well known symptoms, dengue can be mistaken for another viral infection, as they share several features. Therefore, we cannot rely only on mentions of terms such as fever and headache in the tweets. However, previous works (Gomide et al., 2011; Souza et al., 2014) showed a high correlation between the time series of official dengue reports and Twitter data mentioning the keywords *dengue* and *Aedes*. Thus, we set these same two terms to perform a search throughout our collected data, accounting for misspelling and ignoring letter case.

After retrieving all messages based on the predefined keywords, we need to perform a content analysis. That is, we want to classify the messages according to the sentiment expressed in the textual content. The goal of this task is to retain only the messages presenting the largest evidence of a dengue infection, distinguishing them from tweets using the terms in jokes or other not infection related uses.

Our classification was performed in a supervised manner, requiring manually labeled data to train the classifier. In scenarios of disease surveillance, previous studies have already proposed a set of categories for this classification task. We employed the same taxonomy of Chew and Eysenback (2010) and manually labeled a set of 2000 tweets into five sentiment categories. Table 2 shows the sentiment categories, the associated semantics and some instances of real tweets belonging to each class (translated from the original Portuguese texts).

Table 3
Classifier performance: overall accuracy across all five classes, and measures of precision and recall for the personal experience class.

Accuracy	Precision	Recall
0.659907 ± 0.003148	0.730251 ± 0.007777	0.680961 ± 0.005920

We employed the Lazy Associative Classifier (LAC) (Veloso et al., 2007; 2006) to generate a sentiment model from the training data. The classifier uses association rules to assign textual patterns to the predefined categories. These rules have the form $A \rightarrow C$, where the antecedent of the rule A is composed of textual patterns and the consequent C is one of the sentiment categories (e.g., dengue *and* fever $\rightarrow$ personal experience). Each rule represents a vote to the category in the consequent C and the weight of the vote is given by the confidence of the corresponding rule (Agrawal et al., 1993).

In order to assign each message m to one of the categories, we compute a normalized score which estimates the likelihood that a given sentiment category c_i, among the possible values for the consequent C, is being expressed by a message m. This score is given by

$$p(c_i|m) = \frac{\sum_R w(A \rightarrow c_i)}{\sum_C \sum_R w(A \rightarrow C)}$$

where R is the set of rules generated to classify the message m and $w(A \rightarrow c_i)$ is the weight of a generated rule that has c_i as its consequent. Notice that, this approach allows us to assign the same tweet m to more than one category based on its score. For instance, if c_i and c_j present a high score we could say that the tweet expresses the sentiment of both categories. This can be very useful especially when some of the classes have very similar semantics. In our case, we decided to assign each message to the category presenting the highest score.

We performed a preprocessing step in the content to classify the messages. First, we filtered out accent marks and URL's from the text. Also, we created pairs of consecutive words, called bi-grams (Collins, 1996), to enhance the semantics of the textual patterns by providing more context. Finally, some words, called stop-words, were removed. These are words that do not convey much meaning concerning the message content such as articles and prepositions.

In order to assess the performance of our textual content classification, we applied the classifier to the manually labeled dataset. We performed a k-fold cross validation protocol, with $k = 10$. In this evaluation strategy the dataset is partitioned into k folds of roughly equal sample size. Then, $k - 1$ folds are used to train the model and the remaining single fold is held out for testing. The process is repeated k times, therefore using each of the k folds exactly once as the validation data. The result of each fold is then averaged to obtain a single performance estimation. Due to the different proportion among the sentiment categories, we performed a stratified k-fold cross validation, where each of the k folds has approximately the same proportion of class labels. Table 3 shows the mean and standard deviation for classification overall accuracy as well as the precision and recall measures on the personal experience class. All metrics are averaged over 10 runs of the 10-fold cross validation in our labeled dataset to reduce the potential bias of fold selection.

After preprocessing and classifying the messages, we selected those assigned to the *personal experience* category as they present the largest evidence about a dengue infection. These are the red tweets with a hatched shadow in the schematic Fig. 1 and they are named dengue-labeled. The corresponding set of Twitter users who issued such messages are considered the case individuals. The control individuals are those who never issued a dengue-labeled message during the whole period of analysis.

5.4. Building the case and control trajectories

After analyzing the textual content of our geotagged data to create the case and control groups of individuals, we must build the users' corresponding trajectories. Each trajectory is composed by all messages issued by a given user within the period of analysis. More specifically, we are interested in the spatial coordinates associated with each message to trace the individuals movements over the map.

Recall that users belonging to the case group present at least one dengue-labeled message. Therefore, for each individual case we search throughout the dataset to retrieve all other messages issued by the user. All tweets posted by a case individual are considered case tweets, not only those that are labeled *personal experience*. They are connected by red line segments in Fig. 1. Since there is typically a lag of 7–10 days between when a user is infected and when they become symptomatic, we are implicitly considering that the users must have been infected at some point in their daily movement and not necessarily when and where the dengue-labeled messages were sent. In order to avoid highly active users (e.g., bots), we set an upper limit on the total number of messages issued by each user. We adopted a 5-message-per-day threshold, which represents a maximum of 1825 messages per year. The users with total number of messages above this threshold were excluded from the dataset.

The group of control individuals comprises all users who never posted neither a dengue-labeled message nor a message containing any of the keywords used to filter the data in Section 5.3. We introduce this last constraint to potentially reduce noise. All tweets from a given control individual are considered control tweets and they are connected by blue line segments in Fig. 1. We defined the same threshold on the total number of messages per user to exclude highly active individuals in the control group.

As the number of control users is much larger than the number of case individuals, we employ a sampling strategy to select the individuals. To perform the sampling, we stratified the *case* individuals according to the total number of messages in ranges of ten. In each range, we sampled the number of *control* users as 3 times larger than the number of case users. When the number of control users in a given range was not enough to reach the amount of individuals required by our sampling strategy, we select the remaining individuals randomly from the

Table 4
Data summary: #tweets is the total number of tweets issued from the city; #users is the number of unique users; #cases and #ctrls are the number of case and control individuals; #tw_cases and #tw_ctrls are the number of tweets issued by cases and control individuals, respectively.

City	#tweets	#users	#cases	#ctrls	#tw_cases	#tw_ctrls
Campinas	574,226	20,335	90	226	37,313	64,442
Goiânia	566,114	16,849	54	147	15,933	33,750

immediate next range. This sampling approach allows us to obtain case and control groups with a very similar distribution on the number of messages. Table 4 presents a summary of our final dataset for each selected city.

6. Results and discussion

In this section we perform two different analyses. First, we apply both the visit and infection models described in the previous section to the dataset to search for spatial clusters of dengue infection. Next, we perform a comparison between both models and the traditional Bernoulli spatial scan statistics (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995).

6.1. Spatial cluster analysis

As previously mentioned, we selected the two cities reporting the highest incidence of dengue cases in the year 2015 (among the cities with more than 1 million inhabitants) to perform the analysis, as shown in Table 1. In 2014, Brazil faced severe drought conditions that led to a water supply crisis and an increased use of artificial water storage by the population. These artificial sources of standing water served as breeding places for the dengue mosquito and the following year registered a large increase in dengue reports in Brazil. The cities considered in our analysis were deeply affected by the strong surge of dengue.

To run the models for each city, we defined the scanning regions Z by overlaying an axis-aligned rectangular grid to the city. The grid cells are then combined to accommodate regions with different sizes. Also, we set the number of Monte Carlo replicas to build the reference distribution equal to 499 and the significance level equal to $\alpha = 0.05$. Table 5 presents the results.

The visit model detected one significant cluster in Goiânia and no significant clusters in Campinas. The infection model detected four significant clusters in Campinas and two possible clusters (with borderline p-values, $0.05 < p < 0.1$) in Goiânia. We note that in infectious disease surveillance it may be worthwhile to take borderline significant regions into account, depending on the public health resources available for cluster investigation.

As discussed in Section 4, the visit and infection models consider two different conditional probabilities, given by Eqs. (2) and (3), respectively. In this sense, they exploit the data in a different fashion and can be seen as complementary solutions. In fact, they can find different and separate regions in the search process. The visit model searches for regions where $p(Z) > \tilde{p}(Z)$, i.e., it seeks for regions where case individuals are more likely to visit (more precisely, to post at least one tweet while located in that region) than controls. Since it takes into account the binary information of whether or not each individual visited the region at some point during the study period, the visit model is more prone to find larger regions where a high number of case individuals have visited. This effect can be observed, for instance, in the region detected by the visit model in Goiânia, where a large portion of case individuals have issued a tweet. On the other hand, The infection model searches for regions where $r(Z) > r(\bar{Z})$, i.e., it contrasts the risk of being infected inside a given region against the rest of the map. Since the infection model considers the number of times each individual has gone through (tweeted inside) the region, geographically smaller regions with individual cases issuing tweets more times tend to emerge as clusters. This effect can also be observed in Table 5.

Detected regions should be seen only as an approximation to the real geographical clusters (Kulldorff, 2001). For instance, in Fig. 2, we zoom in to the first region detected by the infection model in Campinas, shown in Table 5. This region has 5 case individuals issuing 21 tweets and 4 control individuals posting 16 messages. In order to improve visualization, we introduced a small and uniform jitter to the spatial locations. The first observation is that the detected region is surrounded by other regions with a large number of case individuals, such as the North East and South East areas of the map. We also introduced lines connecting the tweets issued by the same individual. These lines allow us to see that case individuals visiting the detected region also visited the surrounding regions. These surrounding regions can be targeted for surveillance actions. The detected region is located in a non-residential area, being close to two university campuses, parks and one mall. In this sense, assigning individuals only to their residential addresses would hamper the detection of such regions. While we do not have gold standard data available to verify the quality of our methods, we argue that providing a list of suspect high risk regions can greatly benefit surveillance systems and assist public health decision-making regarding preventive actions.

6.2. Alternate model specification using only tweets prior to infection

In the previous section, the analysis was performed using all locations from which each user tweeted during the entire year of 2015. Although it is much more likely that the person got the disease before they issued the dengue-labeled tweet, we still considered the places visited after the dengue-labeled tweet. One reason to follow the above approach is that the Twitter data is sparse, depending on the user's engagement, and it is unlikely that each user tweets from all of the different locations they visit. Thus, the tweet locations from the remainder of the year are also informative as to places where the individual might have been during the infection period. Several studies show that people have very regular movement patterns (Gonzalez et al., 2008) and therefore our analysis used the remaining data to improve our search for riskier regions.

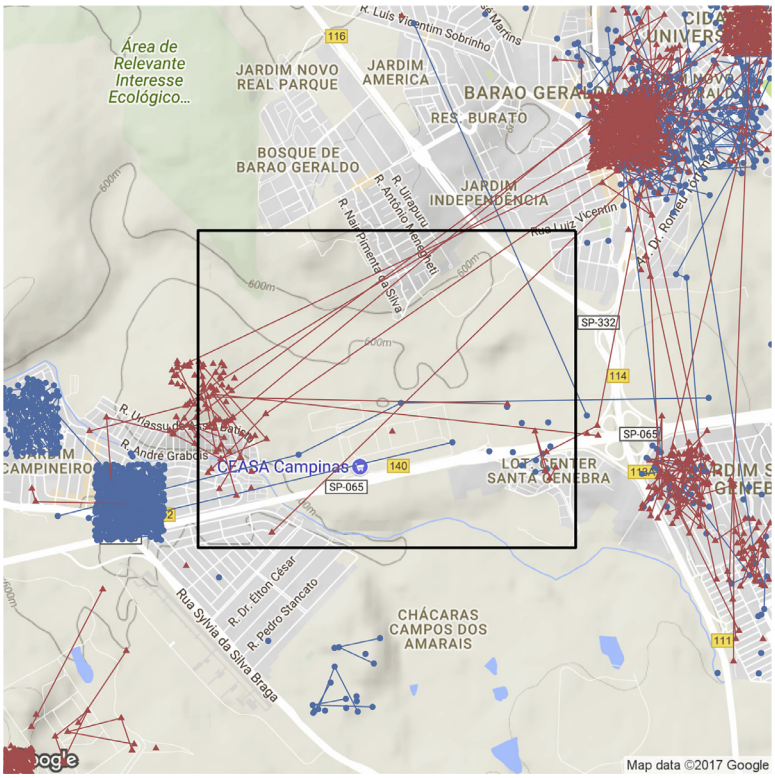


Fig. 2. Zoom in to the first region found in Campinas, identified by the infection model. Detected regions should be seen as a good approximation of the infection places.

In this section, we consider an alternate model specification: for each case individual, we considered only the locations they had been before they issued the dengue-labeled tweet, up to and including the position of the dengue-labeled tweet. If a case individual issued more than one dengue-labeled tweet, we considered only the first one to truncate the data. Although we are not entirely certain of the exact moment the individuals got sick, as they can be mentioning a past case, this alternative approach attempts to capture only the locations visited before the infection manifested. For the control individuals, we adopt a strategy similar to the previous section. We sampled the number of control users as 3 times larger than the number of case users having a number of tweets in the same range of the respective cases. However, this number of tweets is computed in the same time span as the case individuals. This way we are comparing the movements of case and control individuals over the same period. Table 6 shows the details of this new dataset. Notice that, compared to Table 4, this new dataset contains less information about each user's movements due to the more restricted set of tweets.

In order to run both the visit and infection models, we follow the same settings as the previous section. We set the number of Monte Carlo replicas to 499 and the significance level equal to $\alpha = 0.05$. Table 7 presents the results.

Notice that the visit model found a significant cluster in the data from Goiânia. It is noteworthy that the region detected by the visit model in this experiment is very similar to the region found in Section 6.1. Fig. 3 plots both detected regions in a map for comparison. We can see that the regions have a large overlap. This indicates that the visit model was able to find almost the same region using much less data. On the other hand, the infection model did not identify any significant regions for either city in this new dataset. The main explanation is because the infection model depends strongly on the number of times each individual visits a certain region. The truncated dataset is more sparse than the original data used in Section 6.1. Therefore, the infection model has less evidence to identify potentially significant regions.

6.3. Comparison with the spatial scan statistics

In this section, we compare the visit and infection models against the Bernoulli spatial scan statistics (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995). The usual spatial scan assumes that each individual is spatially represented by a single point in the data. Our goal in this section is to demonstrate that this assumption may lead to invalid conclusions in some situations. Thus, we generate two simulated scenarios, described below, and show how directly applying the traditional spatial scan without modification in these settings can result in misleading conclusions. In both scenarios, we use the SaTScan[5] software to run the Bernoulli spatial scan statistics.

[5] https://www.satscan.org/.

Table 5
Visit and infection model results: LL is the log-likelihood of the cluster; $p(Z)$ and $\tilde{p}(Z)$ are the probabilities considered by the visit model; $r(Z)$ and $r(\tilde{Z})$ are the probabilities considered by the infection model; N and M are the respective numbers of case and control individuals inside the zone; N_tweets and M_tweets are the numbers of tweets issued inside the cluster by case and control individuals, respectively.

City	p-value	LL	$p(Z)\,\|\,r(Z)$	$\tilde{p}(Z)\ \|\ r(\tilde{Z})$	N	N_tweets	M	M_tweets
Visit model								
Goiânia	0.01	−135.322	0.044	0.01	48	6352	115	14,600
Infection model								
Campinas	0.006	−695.647	0.07	0.01	5	21	4	16
	0.006	−696.454	0.97	0.01	2	2	0	0
	0.006	−696.499	0.04	0.01	1	49	2	23
	0.006	−696.514	0.97	0.01	2	3	0	0
Goiânia	0.096	−369.431	0.04	0.01	1	36	1	1
	0.088	−367.272	0.22	0.01	1	5	2	2

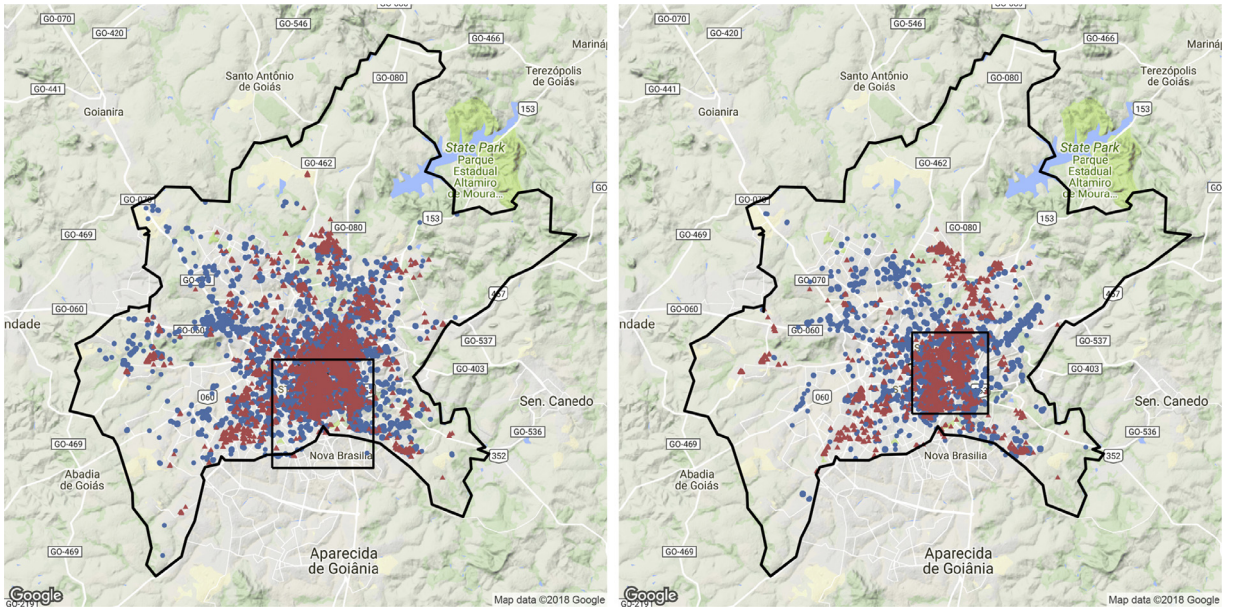


Fig. 3. Results by the visit model in the city of Goiânia. *Left:* using the data from the whole year; *Right*: using the truncated dataset.

Table 6
Total number of tweets from cases and control individuals in the new dataset.

City	#tw_cases	#tw_ctrls
Campinas	16,557	40,925
Goiânia	8951	28,810

Table 7
Results for the truncated dataset: columns are the same as Table 5.

City	p-value	LL	$p(Z)$	$\tilde{p}(Z)$	N	N_tweets	M	M_tweets
Visit model								
Goiânia	0.02	−168.347	0.077	0.01	48	4198	91	11,670

6.3.1. Scenario A

In our first simulation there are 100 control individuals, each one issuing 15 tweets located in space and totaling 1500 positions. These positions are uniformly distributed over the map. The case group has 30 individuals also issuing 15 tweets each, summing up to 450 spatial positions. However, their tweets are distributed differently from the controls. We overlaid a 20×20 grid on the region and selected one cell in this grid to receive 5 tweets from every case individual, totaling 150 tweets in this cell. For each case individual we selected a different, randomly selected cell on the map to receive another 6 tweets belonging to that individual. The remaining 4 tweets per individual are uniformly distributed over the remaining locations on the map. We generated all positions within the boundaries of Goiânia city to make the simulation more realistic.

In order to run the Bernoulli spatial scan, we consider the following approach to preprocess our data: we reduce the set of tweets from each individual user to one single data point in a geographic location by selecting his most common tweeting location. Hence, the total number of data points is equal to the number of distinct individuals in the sample. Each candidate cluster consists of the cells in the 20×20 grid or a connected combination of them. We must then consider the total numbers of case

Table 8
Results for Scenario A. Both the visit and infection models were able to detect the injected cluster. The Bernoulli spatial scan (SaTScan) detects an entirely different region. RR is the relative risk and $\mathbb{E}(N)$ is the expected number of cases.

Algorithm	p-value	LL	$p(Z)\|r(Z)$	$\tilde{p}(Z) \mid r(\tilde{Z})$	N	N_tweets	M	M_tweets
Visit	0.01	−31.417	0.92	0.01	30	150	9	9
Infection	0.01	−21.728	0.41	0.01	30	150	9	9

Algorithm	p-value	LLR	RR	N	$\mathbb{E}(N)$	M
SaTScan	0.00014	14.271	5.23	17	6	9

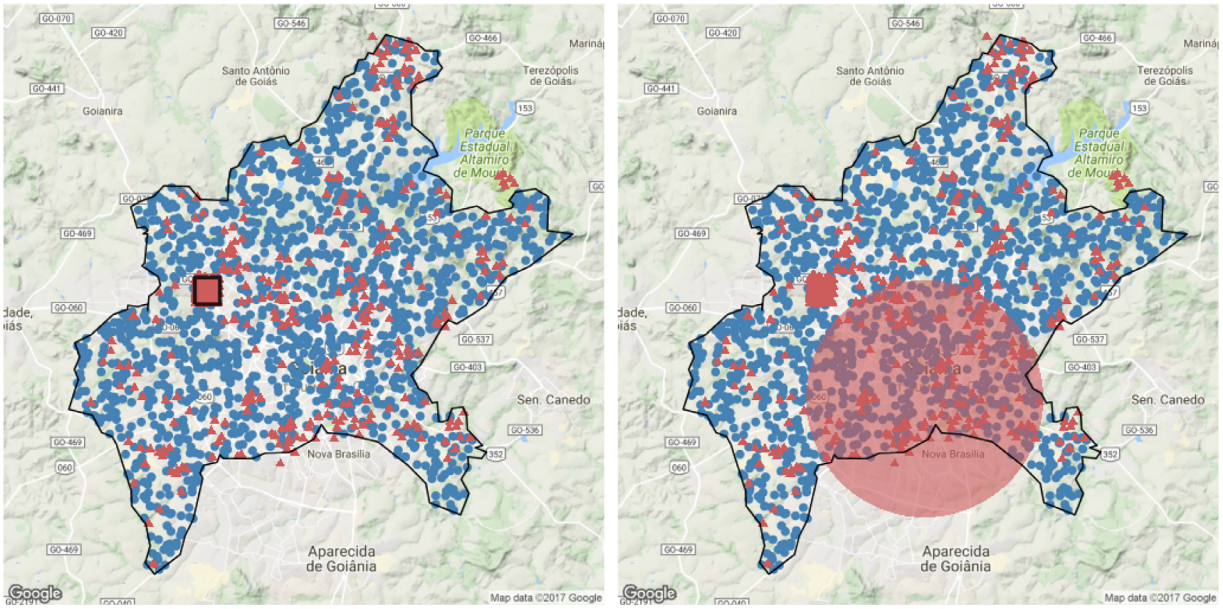


Fig. 4. Left: artificially generated data (Scenario A) and the cluster detected by both the visit and infection models. Right: the cluster detected by SaTScan.

and control individuals for each zone. For SaTScan, we set the maximum size of a cluster to 20% of the population. For the visit and infection models we also set the number of Monte Carlo replicas to 499 and the significance level to $\alpha = 0.05$. Table 8 shows the results.

Intuitively, in this example, we are considering a population of case individuals that live in different places (represented by each individual's most frequent tweeting region) but they share in common another region on the map that they visit frequently, where the infection is assumed to occur. This example illustrates the traditional approach of surveillance systems. Our main goal is to show how this simplifying assumption can create misleading results. From Table 8, we can observe that both the infection and visit models were able to detect the injected cluster. On the other hand, the Bernoulli spatial scan (as implemented by SaTScan) was not able to detect the true cluster as it indeed disappears when each individual's trajectory is reduced to their most frequent location. SaTScan detected another much larger region comprising 20% of the population, its allowed maximum size. Fig. 4 depicts both solutions. The map on the left-hand side shows the generated data along with the region detected by the visit and infection models. On the right-hand side we can see the cluster detected by SaTScan, which does not include the true region. The region detected by SaTScan in this example would not be interesting for public health surveillance since it covers a very large region, lacking the specificity to take immediate actions.

6.3.2. Scenario B

In our second scenario, we artificially generated case and control populations as follows: there are 100 control individuals and 15 positions (representing the tweets) for each individual, totaling 1500 points. These positions are uniformly spread over the map. The cases comprise 31 individuals with 30 of them having 10 positions independently and uniformly distributed on the map. The remaining individual has 150 points concentrated in the same position. One can think of this last individual as only, and frequently, tweeting from his home address. This scenario illustrates one of the challenges when dealing with geolocated social media data: users typically have different levels of engagement in social networks and may present different amounts of information. This simulation is assumed to represent a scenario where no regions of elevated risk are present.

In order to run the Bernoulli spatial scan we consider another possible way of pre-processing our data: we ignore the fact that tweets are produced by individual users and

Table 9
Results for Scenario B. The Bernoulli spatial scan (SaTScan) detects the region with one single individual as extremely significant. Both visit and infection models did not detect any significant clusters.

Algorithm	p-value	LLR	RR	N	$\mathbb{E}(N)$	M
Satscan	$< 10^{-15}$	206.810	5.59	150	36.92	10

simply lump all the tweets together into two sets, a case set and a control set; next, we compute the total number of case and control tweets in each candidate cluster. The candidate clusters, number of Monte Carlo replicas, and α threshold are set as in the previous scenario.

Table 9 shows the results. The visit and infection models did not detect any clusters in the data. Even though one of the regions has a large concentration of tweets, both models are able to take into account the fact that one single individual is responsible for all of the excess tweets and therefore the region should not be considered a true cluster. On the other hand, SaTScan pinpointed this region as highly significant, since it presented a high ratio of case to control tweets. As can be seen in Table 9, the number of expected cases was around 37 while the observed number was 150. Indeed, if we considered the variant presented in Section 6.3.1, SaTScan would ignore this region. However, as discussed above, that variant also has serious drawbacks.

7. Concluding remarks

A major problem in spatial disease surveillance is to locate the spatial clusters of infection risk. The primary difficulty lies in the lack of information about the daily movements of the population at risk. Usually, public health officials have only a single spatial location to associate with each individual, the residence address. Occasionally, there is also a work address. This is not enough to accurately locate the high risk zones at a fine-grained spatial resolution. This may be less important if the data is coarsely aggregated, e.g., by county or state, in which case very few of an individual's tweets may occur outside their area of residence. However, if one is interested in identifying high risk regions within a city, to place each individual in a single position in the map is too coarse.

Social network data offers a unique opportunity to obtain information on the spatial movements of individuals. These data are easily available, in large amount and with almost no delay. Furthermore, we can dynamically extract the disease status as cases and controls of the individuals from the textual content. In this paper, we showed how a publicly available social network, Twitter, can be used to provide such rich information. We described in detail how we collected and processed the data so that they can be used in a disease surveillance system. We also presented two statistical models to search for zones of high infection risk. The models differ because one deals with $\mathbb{P}(A|B)$ while the other with $\mathbb{P}(B|A)$, where A is the event that someone is tweeting from a zone Z and B is the event that the person is a case rather than a control individual.

The stochasticity of location data is not appropriate for the usual spatial cluster detection tools such as the traditional spatial scan statistic approach. Each user is represented by a different number of geographic points and the variability of these numbers is large. We showed how the usual statistical approaches can be easily misled if not extended to account for this special structure.

One limitation of our approach is the self-selected sample nature of our data. A random sample of social network users is not a random sample of the at-risk population. There are multiple biases involved in such a sample (Sloan and Morgan, 2015). The probability of belonging to a given social network is likely to be different according to sex, age, social status and many other attributes that may also be related to the individual's mobility pattern and infection risk. This is a serious objection to the use of social media data and should be carefully considered (Lazer et al., 2014; Pollett et al., 2017). However, we feel that there is merit in developing and using these methods for two reasons. First, in poor regions with lack of information and resources, the suggestion of potential regions of high risk may target a higher proportion of the available resources toward regions with larger probability of being true risk clusters. Second, the population coverage of social networks is expected to continue to expand, resulting in a larger and less biased sample of the population. Additionally, we could imagine using these methods not just on geotagged social media data but on user location data more frequently collected from devices such as cell phones. For example, new initiatives have sampled individuals and, upon their consent, tracked their movement 24/7 as well as measured their disease status (case or control) after some time (Freifeld et al., 2010; Rehman et al., 2016).

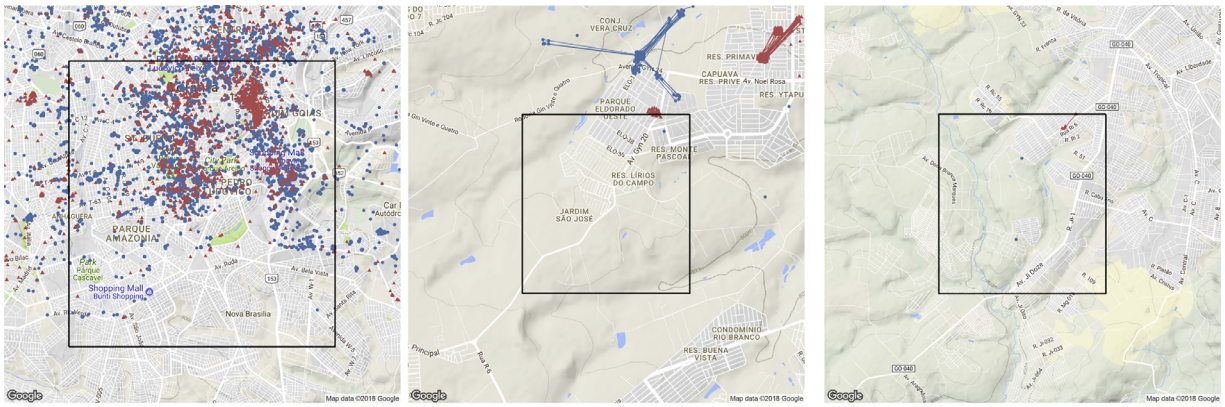
Dengue is just one of many infectious diseases with a well known etiology but a huge amount of uncertain and difficult to obtain parameters that quantify factors such as infected mosquito population, likelihood of being bitten by an infected mosquito, human movement in the mosquito areas, among others. Our methods add to the set of tools that spatial epidemiologists have available to search for spatially localized risk clusters using readily available social network data.

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.

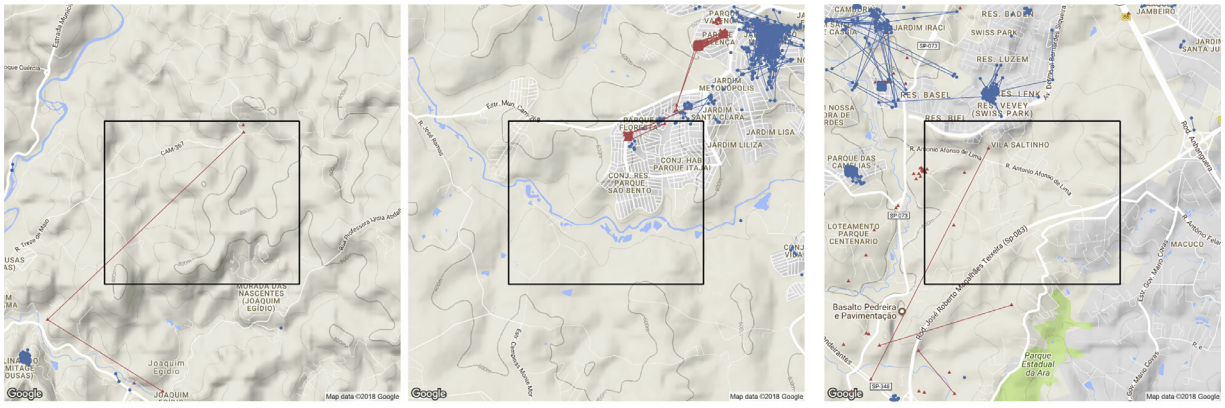
Appendix A. Regions from Table 5

In this appendix, we show a zoom in to the regions in Table 5, complementing Fig. 2. Fig. A.5 depicts the regions from Goiânia, while Fig. A.6 shows the regions detected in Campinas.



(a) Goiânia, region 1. (b) Goiânia, region 2. (c) Goiânia, region 3

Fig. A.5. Zoom in to the regions in Goiânia. Region 1 was identified by the visit model, while regions 2 and 3 were identified by the infection model.



(a) Campinas, region 2 (b) Campinas, region 3 (c) Campinas, region 4

Fig. A.6. Zoom in to the regions in Campinas. All regions were identified by the infection model.

References

Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on management of data, SIGMOD '93; 1993. p. 207–16.

Assunção R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. Stat Med 2006;25(5):723–42.

Bhatt S, et al. The global distribution and burden of dengue. Nature 2013;496.

Bowman LR, Runge-Ranzinger S, McCall P. Assessing the relationship between vector indices and dengue transmission: a systematic review of the evidence. PLoS Negl Trop Dis 2014;8(5):e2848.

Brooker S, Clarke S, Njagi JK, Polack S, Mugo B, Estambale B, Muchiri E, Magnussen P, Cox J. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western kenya. Trop Med Int Health 2004;9(7):757–66.

Chen F, Neill DB. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proceedings of the twentieth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '14. New York, NY, USA: ACM; 2014. p. 1166–75.

Chew C, Eysenback G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. Plos One 2010;5:e14118.

Collins MJ. A new statistical parser based on bigram lexical dependencies. Proceedings of the thirty-fourth annual meeting on association for computational linguistics, ACL '96; 1996. p. 184–91.

Costa MA, Assunção RM, Kulldorff M. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Comput Stat Data Anal 2012;56(6):1771–83.

Cromwell EA, Stoddard ST, Barker CM, Van Rie A, Messer WB, Meshnick SR, Morrison AC, Scott TW. The relationship between entomological indicators of Aedes aegypti abundance and dengue virus infection. PLoS Negl Trop Dis 2017;11(3):e0005429.

Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput Stat Data Anal 2004;45(2):269–86.

Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Ayala Iacucci A, Brownstein JS. Participatory epidemiology: use of mobile phones for community-based health reporting. PLOS Med 2010;7(12):1–5. 12

Gomide J, Veloso A, Meira Jr W, Almeida V, Benevenuto F, Ferraz F, Teixeira M. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Proceedings of the 2011 ACM WebSci conference, 2011.

Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. Nature 2008;453:779–82.

Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N. Childhood Leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. Stat Med 1996;15(7–9):707–15.

Jones SG, Conner W, Song B, Gordon D, Jayakaran A. Comparing spatio-temporal clusters of arthropod-borne infections using administrative medical claims and state reported surveillance data. Spat Spatio-Temporal Epidemiol 2012;3(3):205–13.

Kulldorff M. A spatial scan statistic. Commun Stat – Theory Meth 1997;26(6):1481–96.

Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. J R Stat Soc Ser A (Stat Soc) 2001;164(1):61–72.
Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25(22):3929–43.
Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med 1995;14(8):799–810.
Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science 2014;343(6176):1203–5.
Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clusters as an early warning system for west nile virus activity. Emerg Infect Dis 2003;9(6):641.
Murray NEA, Quam MB, Wilder-Smith A. Epidemiology of dengue: past, present and future prospects. Clin Epidemiol 2013;5:299–309.
Nakaya T, Yano K. Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. Trans GIS 2010;14(3):223–39.
Neill DB. Fast subset scan for spatial pattern detection. J R Stat Soc Ser B (Stat Methodol) 2012;74(2):337–60.
Neill DB, Gorr WL. Detecting and preventing emerging epidemics of crime. Adv Dis Surveill 2007:13.
Neill DB, Moore AW. Rapid detection of significant spatial clusters. Proceedings of the tenth ACM SIGKDD; 2004. p. 256–65.
Paul MJ, Dredze M. You are what you Tweet: analyzing twitter for public health. Proceedings of the 2011 international conference on weblogs and social media (ICWSM); 2011. p. 265–72.
Pollett S, Althouse BM, Forshey B, Rutherford GW, Jarman RG. Internet-based biosurveillance methods for vector-borne diseases: are they novel public health tools or just novelties? PLOS Negl Trop Dis 2017;11(11):1–13.
Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: A good place to detect health conditions. PLOS One 2014;9(1):1–11.
Rehman NA, Kalyanaraman S, Ahmad T, Pervaiz F, Saif U, Subramanian L. Fine-grained dengue forecasting using telephone triage services. Sci Adv 2016;2(7):e1501215.
Shi L, Janeja VP. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). Proceedings of the fifteenth SIGKDD; 2009. p. 767–76.
Sloan L, Morgan J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PLOS One 2015;10(11):1–15. 11
Somanchi S, Neill DB, Parwani AV. Discovering anomalous patterns in large digital pathology images. Stat Med 2018;37(25):3599–615.
Souza RCSNP, Assunção R, Oliveira DM, Brito DEF, Meira Jr W. Infection hot spot mining from social media trajectories. Proceedings of the 2016 ECML/PKDD, 2016.
Souza RCSNP, de Brito DEF, Cardoso RL, Oliveira D JWM, Pappa GL. An evolutionary methodology for handling data scarcity and noise in monitoring real events from social media data. Proceedings of the fourteenth IBERAMIA conference; 2014. p. 295–306.
Stoddard ST, Forshey BM, Morrison AC, Paz-Soldan VA, Vazquez-Prokopec GM, Astete H, Reiner RC, Vilcarromero S, Elder J, Halsey ES, Kochel TJ, Kitron UD, Scott TW. House-to-house human movement drives dengue virus transmission. Proc Natl Acad Sci 2013;110(3):994–9.
Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Paz Soldan V, Kochel TJ, Kitron U, Elder JP, Scott TW. The role of human movement in the transmission of vector-borne pathogens. PLOS Negl Trop Dis 2009;3(7):1–9.
Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 2005;4(1):11.
Veloso A, Meira Jr W, Gonçalves M, Zaki M. Multi-label lazy associative classification. Proceedings of the eleventh European conference on principles and practice of knowledge discovery in databases, ECMLPKDD'07. Springer-Verlag; 2007. p. 605–12.
Veloso A, Meira Jr W, Zaki MJ. Lazy associative classification. Proceedings of the 2006 international conference on data mining; 2006. p. 645–54.