Detecting Spatial Clusters of Infection Risk with Geo-Located Social Media Data

Roberto C.S.N.P. Souza
Supervised by Prof. Wagner Meira Jr. and by Prof. Renato Assunção
Department of Computer Science – Universidade Federal de Minas Gerais
nalon@dcc.ufmg.br

ABSTRACT

Spatial health surveillance systems typically identify high risk regions based only on the residence address of diseased individuals. Geo-located social media data offers a unique opportunity to obtain information on the spatial movements of individuals as well as their disease status. This can be a rich source of information to identify high risk places even in regions where no one lives such as parks and entertainment zones. We develop two models and their respective algorithms to deal with this challenging problem. We demonstrate the applicability and effectiveness of our proposed methods by applying them to a collection of geo-tagged Twitter messages coming from Brazil. In particular, we target the identification of spatial clusters associated with Dengue, a vector-borne disease that affects millions of people in Brazil annually, and billions worldwide, to show the usefulness of our methods for disease surveillance.

1. INTRODUCTION

The current fast increasing popularity of devices equipped with location sensors offers unprecedented possibilities for data mining research. The deluge of geo-located data daily generated across several sources (e.g. mobile phones, connected vehicles) allows us to answer many questions regarding the behavior of targeted populations with a timing and precision not possible in the last decade.

For instance, 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 geo-located 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 research, we give a contribution in such direction. We exploit geo-located social media data to detect spatial clusters of disease infection. Our main goal is to contrast observed mobility patterns for diseased and non-diseased individuals in order to detect localized regions where likelihood of being infected by a given disease is higher than in the rest of the map. Identifying places where people have higher risk of being infected may be key to surveillance actions. In summary, our contributions are as follows: (i) We introduce the problem of detecting spatial clusters of infection risk from mobility data; (ii) We propose two novel models, and their respective algorithms, for the discovery of spatial clusters of infection risk; (iii) We propose an extraction and modeling strategy of geo-located social media data to the proposed problem; (iv) We present experimental results to illustrate our approach in action by applying our algorithms to a collection of geo-located Twitter data from Brazil.

2. PROBLEM DESCRIPTION

We use Figure 1 to explain our problem. Each individual is indexed by a number i and has a set of n_i spatial positions. The spatial points are given by each lat-long coordinate pair embedded in a geo-located social media post. The positions from a single person are connected by line segments in Figure 1, representing her movements. The individuals are additionally labeled by two colors according to their status: disease cases (in red) or controls (in blue). The cases are those individuals who mentioned a personal experience with the disease in at least one message (further details in Section 4). The messages which have mentioned personal experience with the disease are marked with a hatched shadow in Figure 1. The figure also shows a spatial region Z where the risk of becoming a case might be higher than in the rest of the region.

Our main goal is then to scan the map varying the position, shape, and size of the candidate regions, looking for the region $\hat{Z}$ that most likely is a higher risk area. After finding this most likely hot spot $\hat{Z}$, we want to calculate its probability of occurrence to evaluate whether there is enough evidence to call it a real cluster.

Several challenges emerge in this problem: (i) Each i-th individual is not associated with a single location, as in the usual spatial cluster detection task [2, 5, 4], but rather with a series of n_i successive positions on the map; (ii) The number n_i of positions of each individual is quite variable, depending on her social media usage. Clearly, the locations can

Proceedings of the VLDB 2018 PhD Workshop, August 27, 2018. Rio de Janeiro, Brazil.

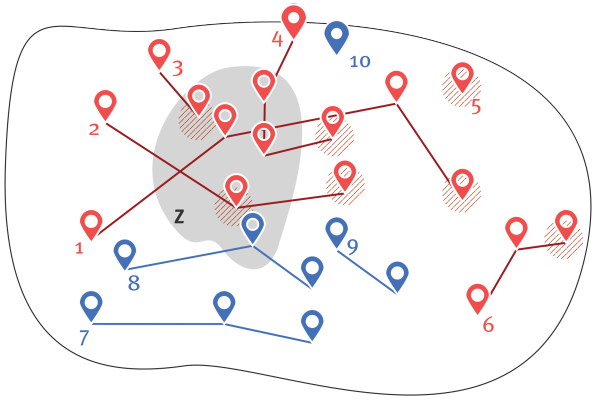

Figure 1: Schematic drawing of a potential infection risk region (shaded area) and the individuals movements of cases (red) and controls (blue).

not be put on the map ignoring the different contribution of each individual, otherwise, an extreme individual would dominate the analysis; (iii) The positions of the disease-labeled messages (indicating personal experience) are not necessarily those where the infection risk is higher. Indeed, our assumption is that the individual entire mobility pattern (and not a single position) will be informative of the high risk areas.

3. DETECTING SPATIAL CLUSTERS

We adopted a case-control framework, where the data consist of locations, within a specified geographical region, of all known cases of a particular disease, and of a random sample of controls drawn from the population at risk. We labeled the individuals such that the first N of them are the cases and the last M are the controls.

In our analysis, the key innovation is that the input is a series of locations rather than a single location for each individual. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n_i})$ be the point events associated with the n_i messages issued by the i-th individual, $i = 1, \ldots, N+M$. Each $x_{i,k}$ represents the geographical message location such as a lat-long coordinate pair. For the cases $i = 1, \ldots, N$, at least one message in $\mathbf{x}_i$ refers to a personal experience with the disease and it is denoted disease-labeled message. Typically, there will be a small percentage of disease-labeled messages for each individual. None of the control individual messages are disease-labeled.

Let $\mathcal{Z}$ be a (large) set of geographical regions that are candidates to be spatial clusters. There are potentially infinite regions in $\mathcal{Z}$ and they cover the entire region under analysis. By varying $Z \in \mathcal{Z}$ we scan the map looking for the zone $\hat{Z}$ that most likely is a higher risk area. After finding this most likely hot spot $\hat{Z}$, we calculate its likelihood to evaluate whether there is enough evidence to identify it as a real cluster.

The multiple number of 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 call *Visit Model* and *Infection Model* [6].

3.1 Visit Model

Let $V_{i,z}$ be the random number of messages in Z among the n_i total number of messages issued by the i-th individual. Use $\mathbb{1}[A]$ to represent the indicator random variable that the event A occurs. Hence $\mathbb{1}[V_{i,z} \geq 1]$ is the binary random variable indicating whether the i-th individual ever issued a message inside the candidate zone Z. These random variables can be assumed independent, but they are not identically distributed as the success probability depends on the number n_i of messages issued by each individual. Denote by $p = p(Z)$ the probability that, giving that a case individual is tweeting, she does it from within Z. Let $\bar{p} = \bar{p}(Z)$ be the similar probability for a control individual. We are interested in zones where $p(Z) > \bar{p}(Z)$.

For a user who is a case, we have $\mathbb{P}(V_{i,z} \geq 1)$ equals to $1-(1-p)^{n_i}$ and, for a control user, it is equal to $1-(1-\bar{p})^{n_i}$. Considering a fixed zone Z, the visit model likelihood is given by

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

where we simplified the expression by dropping the zone Z from $p(Z)$ and $\bar{p}(Z)$ writing simply p and $\bar{p}$.

3.2 Infection Model

We will estimate the probability that someone issues a disease-labeled message (becomes a case) given that she visited k times the region Z. Let $r = r(Z)$ be the infection risk inside the candidate cluster and $\bar{r} = r(\bar{Z})$ the infection risk in $\bar{Z}$, the region outside Z. We are interested in zones Z where $r(Z) > r(\bar{Z})$.

Let I_i be the binary indicator that the individual i is a case. We assume that these binary random variables are independent. They are not identically distributed since their probability of $I_i = 1$ depends on the number of visits $V_{i,z}$ by the i-th individual to the zone Z. We have $\mathbb{P}(I_i = 1 | V_{i,z} = k_i) = 1 - \mathbb{P}(I_i = 0 | V_{i,z} = k_i) = 1 - (1-r)^{k_i} (1-\bar{r})^{n_i - k_i} = \pi(k_i, r, \bar{r})$. Therefore, the likelihood of the pattern of cases and controls is given by

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

3.3 Evaluating the Data Evidence

Recall that $\mathcal{Z}$ is the set of candidate zones to be scanned. The test statistic we adopt for the Visit Model is

$$T_1 = L_1(\hat{Z}, \hat{p}, \hat{\bar{p}}) = \sup_{\substack{Z \in \mathcal{Z} \\ \hat{p}(\hat{Z}) > \hat{\bar{p}}(\hat{Z})}} L_1(\hat{Z}, \hat{p}(\hat{Z}), \hat{\bar{p}}(\hat{Z})) \qquad (1)$$

and an analogous formula defines T_2 for the Infection Model. In order to verify its statistical significance, we must use Monte Carlo simulation to obtain the null hypothesis distribution of T_1 and T_2 as the exact or asymptotic analytic calculation is not feasible. The null hypothesis is given by either $H_0 : p = \bar{p}$ or $H_0 : r = \bar{r}$ for all $Z \in \mathcal{Z}$ for the Visit Model and the Infection Model, respectively.

The Monte Carlo distribution is determined by randomly permuting the labels of cases and controls among all individuals. Using this pseudo dataset, we proceed the entire scan over all $Z \in \mathcal{Z}$ to obtain a pseudo value for T_1 and T_2. As this will be replicated several times, we call these values $T_1^{(1)}$

and $T_2^{(1)}$. We then select another random permutation of the labels, scan the zones and find $T_1^{(2)}$ and $T_2^{(2)}$. Independently, we repeat this procedure a large number $B-1$ of times generating a set of pseudo values plus the values calculated with the actually observed dataset: $T_1, T_1^{(1)}, T_1^{(2)}, \ldots, T_1^{(B-1)}$ and $T_2, T_2^{(1)}, T_2^{(2)}, \ldots, T_2^{(B-1)}$. Under the null hypothesis, these values are independent and identically distributed. Therefore, the rank of the real observed statistics T_1 and T_2 are uniformly distributed on the integers $1, \ldots, B$. This implies that an exact p-value for the null hypothesis of visit model is given by

$$p_1 = \frac{1}{B}(1 + \#\{T_1^{(k)} \geq T_1, k = 1, \ldots, B-1\})$$

and an analogous formula defines p_2 for the Infection Model. The test is significant at the level $\alpha \in (0,1)$ if $p_m < \alpha$. When either test is significant, the most likely zone is given by the corresponding maximizing argument $\hat{Z}$ in (1).

We also identify secondary clusters, zones with highly significant p-values, which do not intersect with the most likely zone $\hat{Z}$. The non-intersecting restriction is necessary because, if one zone $\hat{Z}$ is the most anomalous in $\mathcal{Z}$, many other sets in $\mathcal{Z}$ that are only slightly different from $\hat{Z}$ will produce very similar likelihood numbers. These zones should be ignored since the most anomalous among them has already been pinpointed. Among the non-intersection zones, we look for those whose p-value p_m is smaller than α where the p-values are calculated as described above.

4. EXPERIMENTAL ANALYSIS

In this section, we apply both models to search for spatial clusters of Dengue infection using Twitter data.

Dengue Overview: Dengue is an emerging mosquito-borne viral disease with estimated 100 million global infections per year [1]. Brazil reports more cases 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 World Health Organization (300 cases). Dengue has a huge amount of uncertain and hard to obtain (if feasible) parameters driving the disease. Human mobility is one of the key factors, especially due to the mosquito day-biting habit [8]. Thus, 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.

Dataset: Our geo-located data were collected through the Twitter Streaming API[2]. The collection period goes from January 1st, 2015 to December 31th, 2015 during which we were able to crawl a total of 106,784,441 Twitter messages. We set a geographic boundary box covering the Brazilian territory and consequently all collected tweets are geo-tagged with lat/long GPS coordinates. Based on the geographic coordinates, we assigned each tweet to a valid municipality to perform a city-level analysis.

Content Filtering and Analysis: Individuals are labeled as cases or controls based on the content of their tweets. In order to find individuals presenting a dengue infection episode, we follow [7] and search for all the tweets presenting the keywords *dengue* and *Aedes*. These messages are then classified into one of five categories: personal experience, information, campaign, opinion and irony/sarcasm. After classifying the messages, the group of case individuals are defined as those users who issued at least one tweet assigned to the *personal experience* category. The control individuals group is composed by the remaining users. Notice that, the individuals' mobility patterns are composed by the locations of all messages they issued in the period.

In order to run the algorithms, the zones Z are defined by overlaying different grids on the map and each grid cell corresponds to a zone to be scanned. The size of the grid cells vary in order to accommodate risk zones that present different characteristics. We set the number of Monte Carlo replicas to $B - 1 = 999$ and define the significance level as $\alpha = 0.05$. We present the results for 4 Brazilian cities in Table 1, Goiânia (GOI), Limeira (LIM), São José dos Campos (SJC) e Sorocaba (SOR).

Table 1: Results for Visit and Infection models. LL is the log-likelihood; $r \mid p$ and $\bar{r} \mid \bar{p}$) are probabilities considered by the models; p-v is the value; N and M are the number of case/control individuals in the zone; N_k_i and M_k_i are the number of messages issued inside the zone by case/control individuals.

City	LL	$r \mid p$	$\bar{r} \mid \bar{p}$	p-v	N	N_k_i	M	M_k_i
Visit Model								
GOI	-135.32	0.04	0.01	0.01	48	6352	115	14600
LIM	-89.52	0.04	0.01	0.019	43	5655	80	7940
Infection Model								
	-198.51	0.48	0.01	0.014	5	11	1	1
LIM	-200.16	0.07	0.01	0.02	4	8	3	10
	-200.35	0.07	0.01	0.02	3	97	7	9
SJC	-427.44	0.14	0.01	0.055	5	28	2	4
SOR	-446.95	0.04	0.01	0.002	3	150	8	16

Notice that our models were able to find spatial clusters in 3 cities that faced strong surges of Dengue during the year of 2015. This is a good indication of the usefulness of the algorithms for disease surveillance. Figure 2 depicts the zones found by each model in the corresponding cities. Notice that in the city of Limeira the models identified different regions within the same city. These results also point out the complementarity of the models, so that they may be used together towards establishing two different levels of surveillance.

5. RELATED WORK

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 [2, 3, 4], 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

[1] http://www.paho.org/data/index.php
[2] https://dev.twitter.com/streaming/overview

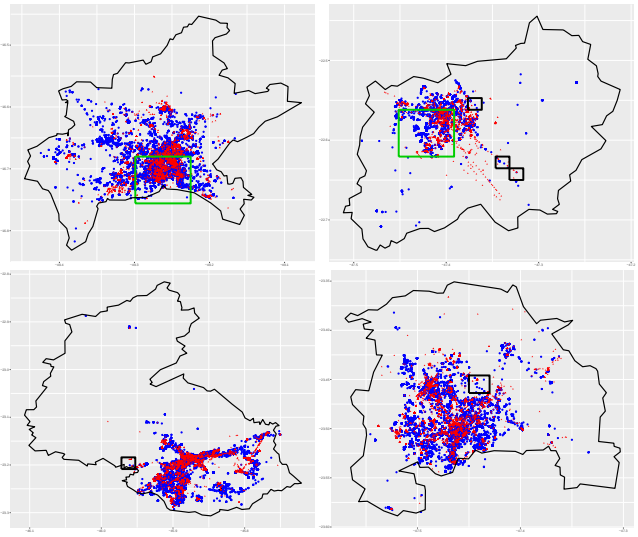

Figure 2: Maps of the cities with the hot spots found by both models. The cities are Goiânia, Limeira, São José dos Campos and Sorocaba. The green and black squares depict the zones found by the Visit and Infection models respectively. We also display the case and control individuals spatial points as red and blue points, respectively.

provide meaningful statistical measures to evaluate the significance of detected clusters. A major application of spatial cluster detection methods is the detection of disease clusters to suggest risk factors, to focus preventive efforts, and for outbreak monitoring [3, 4]. However, they have also been applied to other tasks, such as the identification of hot spots zones based on the locations of traffic accidents [5].

In all this large body of work there has been one invariant aspect of the spatial characterization of the individual input data: there is only one spatial position associated with each one of them. Let it be a pixel (as in a medical image) or a random spatial event (such as an accident location), they have one single spatial location associated, either it be a case or a control individual. 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, his residential and working place addresses. The recent availability of spatial data offers a unique opportunity and the existing data mining techniques for spatial cluster detection fail to address this new setting as they require a single location to each individual under analysis. Our proposal generalizes this approach by considering the individuals' mobility patterns instead of a single point.

6. CONCLUSIONS

Geo-located social media data offers a unique opportunity to obtain information on the spatial movements of people. These data are easily available, in large amount and with almost no delay. Furthermore, we can extract the disease status as cases and controls of the individuals from the textual content. The stochasticity of location data is not appropriate for the usual spatial cluster detection tools such as the traditional spatial scan statistic approach [2, 4]. Each user is represented by a different number of geographic points and the variability of these numbers is large.

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 risk population. There are several biases involved in such a sample. 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 geo-tagged 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.

7. ACKNOWLEDGMENTS

We would like to thank FAPEMIG, CNPq and CAPES for their financial support. This work was also partially funded by projects InWeb, MASWeb, EUBra-BIGSEA, INCT-Cyber, ATMOSPHERE and by the Google Research Awards for Latin America program.

8. REFERENCES

[1] S. Bhatt et al. The global distribution and burden of dengue. *Nature*, 496, 2013.

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

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

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

[5] L. Shi and V. P. Janeja. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). In *Proc. of SIGKDD*, 2009.

[6] R. C. S. N. P. Souza, R. Assunção, D. M. Oliveira, D. E. F. Brito, and W. Meira Jr. Infection hot spot mining from social media trajectories. In *Proc. of the ECML/PKDD*, 2016.

[7] R. C. S. N. P. Souza et al. An evolutionary methodology for handling data scarcity and noise in monitoring real events from social media data. In *14th IBERAMIA Conference*, pages 295–306, 2014.

[8] S. T. Stoddard et al. House-to-house human movement drives dengue virus transmission. *PNAS*, 110(3):994–999, 2013.