Identifying High-Risk Areas for Dengue Infection Using Mobility Patterns on Twitter

Roberto C. Souza[1], Daniel B. Neill[2], Renato M. Assunção[1], Wagner Meira[1]

[1]Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, [2]New York University, New York, New York, United States

Objective

We develop new spatial scan models that use individuals' movement data, rather than a single location per individual, in order to identify areas with a high relative risk of infection by dengue disease.

Introduction

Traditionally, surveillance systems for dengue and other infectious diseases locate each individual case by home address, aggregate these locations to small areas, and monitor the number of cases in each area over time. However, human mobility plays a key role in dengue transmission, especially due to the mosquito day-biting habit [1], and relying solely on individuals' residential address as a proxy for dengue infection ignores a multitude of exposures that individuals are subjected to during their daily routines. Residence locations may be a poor indicator of the actual regions where humans and infected vectors tend to interact more, and hence, provide little information for dengue prevention. The increasing availability of geolocated data in online platforms such as Twitter offers a unique opportunity: in addition to identifying diseased individuals based on the textual content, we can also follow them in time and space as they move on the map and model their movement patterns. Comparing the observed mobility patterns for case and control individuals can provide relevant information to detect localized regions with higher risk of dengue infection. Incorporating the mobility of individuals into risk modeling requires the development of new spatial 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. We propose new spatial scan models and exploit geo-located data from Twitter to detect geographic clusters of dengue infection risk.

Methods

As the spatial tracking of a large sample of infected and non-infected individuals is expensive and raises serious privacy issues, we instead analyze geo- located Twitter data (tweets), which is readily and publicly available. We identify "infected" individuals (cases) as those individuals who have at least one tweet classified as a current, personal experience with dengue. We note that, because of the incubation period and recovery time, infected Twitter users are likely to mention dengue in their tweets days after they are infected, and usually not at the location where the exposure (mosquito bite) occurred. Once we have identified cases and controls based on the textual content of the messages, we then compare the mobility patterns of the two groups. The key aspect of our method is that the input is a series of locations rather than a single location, such as the residence address, for each individual. The number of positions n_i composing each mobility pattern can vary substantially between individuals i, and thus simple approaches like counting the total numbers of case and control tweets per location would be biased and inaccurate; moreover, individuals with larger numbers of tweets may be more likely to be identified as a case. Nevertheless, our assumption is that the entire mobility patterns will be informative of the riskier areas if we compare the spatial patterns from infected and non-infected individuals.

We have developed two new spatial scan methods (unconditional and conditional spatial logistic models) which correctly 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; the estimate is biased by individuals' propensity to tweet in different locations, but is expected to capture the large amounts of time spent at frequently visited locations. Our unconditional model controls the variable contribution of each individual through a non-parametric estimation of the odds of being a case and has a semi-parametric logistic specification. When estimating the previous offset becomes a complex task, we propose a case-control matching strategy in the conditional model to control for the number of tweets n_i. Based on the subset scan approach [2], 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 statistic over subsets of the data.

Results

We demonstrate the detection of high-risk clusters for dengue infection using Twitter data we collected in Brazil during the year of 2015, when a strong surge of dengue hit several cities. We apply our method to the cities with highest number of case individuals. There are many points of interest, such as hospitals and parks, inside the detected regions. As those places are non-residential, standard approaches would fail to consider them as potential infection places in the event of a spike in the number of cases. Figure 1 shows the detected regions in the city of Campinas, Brazil. Synthetic and real-world evaluation results demonstrate that our methods work better than either just mapping each individual to their most frequent location (which is a proxy for home address) and running a traditional spatial scan, or scanning using tweet volume as an input.

Conclusions

Identifying places where people have higher risk of being infected, rather than focusing on residential address locations, may be key to surveillance for vector-borne diseases such as malaria and dengue, allowing public health officials to focus mitigation actions. The stochasticity of location data is not appropriate for typical spatial cluster detection tools such as the traditional spatial scan statistic [3]. 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. 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. Our methods add to the set of tools that spatial epidemiologists have available to search for spatially localized risk clusters using readily available Twitter data. We expect that our method will also be useful to other public health surveillance problems where movement data can bring relevant information.

Acknowledgement

This work was partially funded by FAPEMIG, CNPq and CAPES and also by the projects InWeb, MASWeb, EUBra-BIGSEA, INCT-Cyber, ATMOSPHERE and by the Google Research Awards for Latin America program.

References

1. Stoddard ST, et al. 2009. The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis.* 3(7), e481. PubMed https://doi.org/10.1371/journal.pntd.0000481
2. Neill DB. 2012. Fast subset scan for spatial pattern detection. *J R Stat Soc B*. 74(2), 337-60. https://doi.org/10.1111/j.1467-9868.2011.01014.x
3. Kulldorff M. 1997. A spatial scan statistic. *Commun Stat Theory Methods*. 26(6), 1481-96. https://doi.org/10.1080/03610929708831995

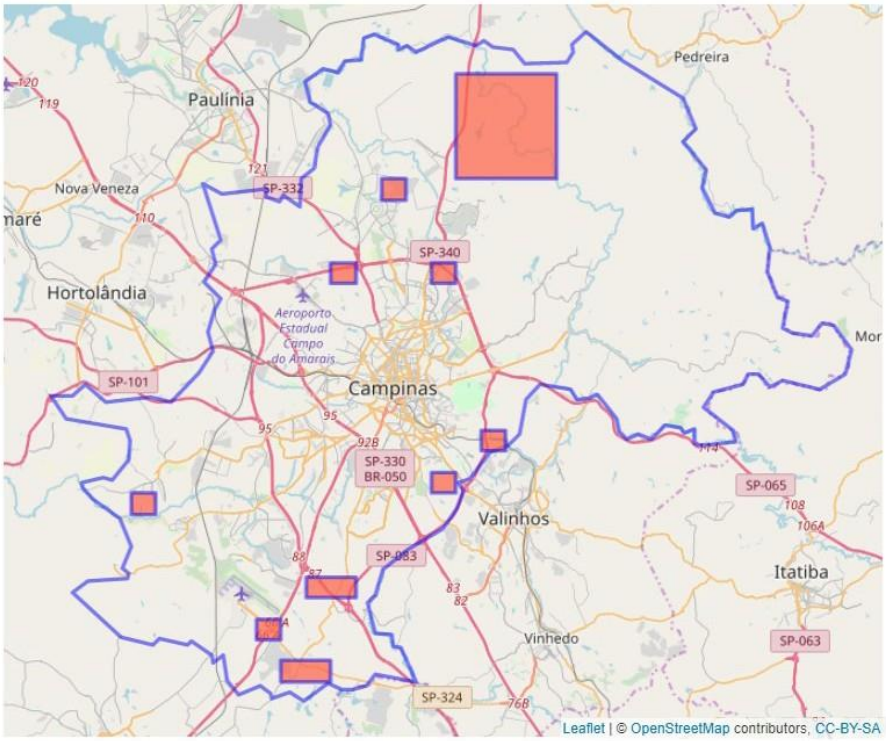

Figure 1: Detected regions in the city of Campinas, Brazil.