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

Roberto C.S.N.P. Souza[1], **Daniel B. Neill**[2], Renato Assunção[1], Wagner Meira Jr.[1]

[1] Universidade Federal de Minas Gerais

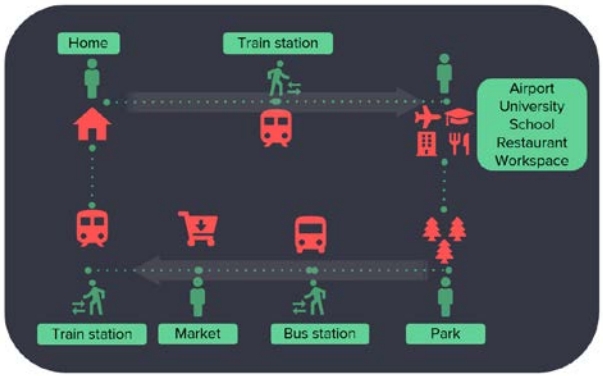
[2] Machine Learning for Good Lab, New York University

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

NYU
ML for Good

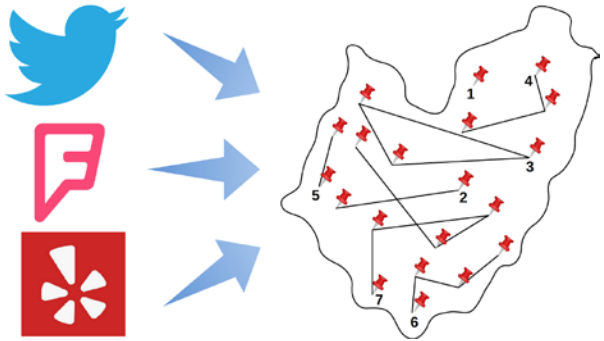
Dengue surveillance

- Traditional dengue surveillance systems locate each individual by **home address**, and monitor the # of cases in each area over time.
- However, home location may be a poor indicator of **exposure risk**. *Aedes aegypti* mosquitos bite during the day; most exposures happen away from home.
- **Human mobility** plays a key role in dengue transmission!



Dengue surveillance and social media

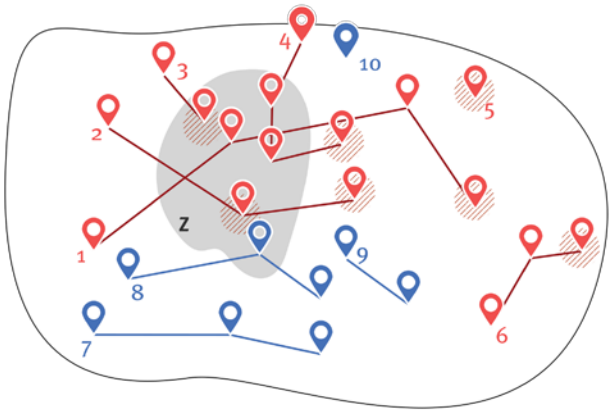
There is an increasing availability of **geolocated data** in online platforms, such as **Twitter**. More than just locating an individual at a given time, this provides an estimate of typical **mobility patterns**.



Based on this data, we can **identify** diseased individuals using textual content of their messages and **follow** them in time and space as they move on the map.

Problem definition

- We wish to identify a spatial region Z where the **risk** of being infected by dengue is higher than in the rest of the map.
- Each individual is represented by a set of positions describing their **movements**.



Case individuals have at least one tweet classified as a current, personal experience with dengue.

Control individuals do not experience dengue in the given time period.

Challenges

Dataset challenges:

- 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 available but only gives snapshots of the user's position.

Methodological challenges:

- The **number of positions** n_i (tweets) composing each mobility pattern can vary substantially between individuals i.
- Because of the **incubation period** and recovery time, infected users are likely to mention dengue in their tweets days after infection, and usually not at the location where the exposure occurred.

Two new spatial scan models

- We developed two novel spatial scan methods that take mobility patterns into account, the **unconditional spatial logistic model** and the **conditional spatial logistic model**.
- Both models use the proportion of an individual's tweets as a rough estimate of the proportion of time spent in each location.
- This estimate is biased by individuals' propensity to tweet in different locations, but is nevertheless expected to capture the large amounts of time spent in frequently visited locations.

Two new spatial scan models

- Both models assume that each individual i's risk (log-odds of being a dengue case rather than a control) is increased by some constant β times the proportion of time $p(Z)_i$ spent in the high-risk region Z.

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

- One complication is that individuals with more tweets n_i are more likely to tweet about a personal experience with dengue, and thus more likely to be identified as cases, regardless of mobility pattern.

- The **unconditional model** fits the function $g(ni)$ to account for this, while the **conditional model** instead matches cases and controls with similar n_i so that their $g(ni)$ values cancel.

Two new spatial scan models

- As is usual for spatial scan, we identify the region Z that maximizes the generalized **log-likelihood ratio** (LLR) test statistic:

$$LLR(Z) = \max_{\beta} \log \frac{\Pr(Data \mid H_1(Z), \beta, g(n_i))}{\Pr(Data \mid H_0, g(n_i))}$$

- We compute the statistical significance (p-value) by **randomization testing**, comparing $LLR(Z)$ to the highest-scoring regions found in simulated datasets generated under the null hypothesis $\beta = 0$.

Case Study: Dengue in Brazil

Dengue in Brazil

- Brazil reports more cases of dengue than any other country.
- In 2015 the Brazilian Ministry of Health reported **~1.6 million cases** of dengue.
- **839 deaths** were confirmed to be caused by dengue in the same year.
- More than **US $300 million** was spent to fight the disease in 2015.

Government campaign:
Everyone against dengue!

Dataset

Data Acquisition

- The data were collected from Twitter, using the Streaming API.
- All tweets are geo-tagged with lat-long coordinates.
- More than 100 million geo-tagged Twitter messages from January 1st to December 31st, 2015.

Data Preprocessing

- City-level analysis: In Brazil each city hall is in charge of the decision-making regarding dengue surveillance actions.
- Data were filtered by latitude/longitude.
- We selected the city of **Sorocaba** as a case study.

Case study: City of Sorocaba, Brazil

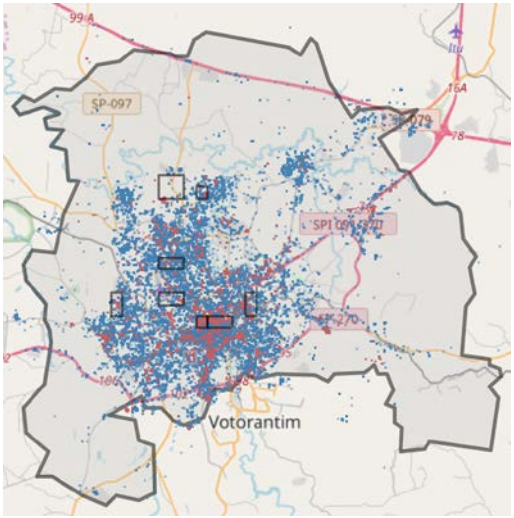
- The city of Sorocaba is located in the Southeast region of Brazil (close to São Paulo) with a total population of ~650K inhabitants.
- It was one of the most affected cities in the strong dengue surge in Brazil in 2015, reporting less than 400 dengue cases in 2014 but more than **50 thousand cases** in 2015.
- Likely explanation: the huge drought that affected Brazil in 2014.

  Without abundant water, the mosquito has difficulty breeding, leading to a low number of cases. However, in face of a drought people tend to store water in improper places, creating a large number of mosquito **breeding sites** which can lead to severe outbreaks.

Detected risk clusters for Sorocaba

Regions detected in Sorocaba (unconditional model).

LLR	β	p-value	#cas	#ctl	#tw_cas	#tw_ctl
17.825	0.184	0.005	22	30	2190	122
17.755	2.719	0.005	9	4	200	5
15.342	0.150	0.005	13	11	331	108
11.384	1.289	0.005	5	4	503	8
10.880	0.408	0.005	14	31	70	76
10.674	0.144	0.005	30	40	178	278
8.006	2.338	0.045	4	3	115	3
7.748	9.710	0.050	4	0	8	0
7.503	0.649	0.050	8	9	69	27



Detected regions include several **non-residential places**, including hospitals, parks, college campuses, etc.

Standard approaches would not be able to detect such regions!

Conclusions

- Identifying the highest-risk places for infection would benefit dengue surveillance by targeting **prevention** and **mitigation** actions where they are most needed.
- **Twitter** and other social media offer a unique opportunity to obtain information on the spatial movements of individuals, and can provide **useful input** for surveillance of dengue and other diseases.
- The challenges of mobility data require **new methods** that can cope with this type of data in a principled way.
- Our methods add to the **set of tools** that public health has available to search for **spatially localized risk clusters** using readily available Twitter data, or other geo-located data sources.

References

- R. C. Souza, R. M. Assunção, D. M. de Oliveira, D. E. de Brito, and W. Meira Jr. **Infection hot spot mining from social media trajectories.** In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 739–755. Springer, 2016.
- R. C. Souza, R. M. Assunção, D. B. Neill, L. G. Silva, and W. Meira Jr. **Spatial risk modeling for infectious disease surveillance using population movement data.** NeurIPS Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 2018.
- R. C. Souza, R. M. Assunção, D. M. de Oliveira, D. B. Neill, and W. Meira Jr. **Where did I get dengue? Detecting spatial clusters of infection risk with social network data.** Spatial and Spatio-temporal Epidemiology, 2018.

Thanks for listening!

More details on my web site:
http://www.cs.nyu.edu/~neill

Or e-mail me at:
daniel.neill@nyu.edu