

Enhancement of Epidemiological Models for Dengue Fever Based on Twitter Data

Julio Albinati, Wagner Meira Jr., Gisele L. Pappa

Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
jalbinati,meira,glpappa@dcc.ufmg.br

Mauro Teixeira, Cecilia Marques-Toledo

Department of Biochemistry and Immunology
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
mmtex@icb.ufmg.br,cecilia.a.marques@gmail.com

ABSTRACT

Epidemiological early warning systems for dengue fever rely on up-to-date epidemiological data to forecast future incidence. However, epidemiological data typically requires time to be available, due to the application of time-consuming laboratorial tests. This implies that epidemiological models need to issue predictions with larger antecedence, making their task even more difficult. On the other hand, online platforms, such as Twitter or Google, allow us to obtain samples of users' interaction in near real-time and can be used as sensors to monitor current incidence. In this work, we propose a framework to exploit online data sources to mitigate the lack of up-to-date epidemiological data by obtaining estimates of current incidence, which are then explored by traditional epidemiological models. We show that the proposed framework obtains more accurate predictions than alternative approaches, with statistically better results for delays greater or equal to 4 weeks.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → *Expert systems*; *Social networks*; • **Computing methodologies** → *Gaussian processes*;

KEYWORDS

dengue fever, Twitter, Gaussian processes

ACM Reference format:

Julio Albinati, Wagner Meira Jr., Gisele L. Pappa and Mauro Teixeira, Cecilia Marques-Toledo. 2017. Enhancement of Epidemiological Models for Dengue Fever Based on Twitter Data. In *Proceedings of DH '17, London, United Kingdom, July 02-05, 2017*, 10 pages.
<https://doi.org/http://dx.doi.org/10.1145/3079452.3079464>

1 INTRODUCTION

Dengue fever is a mosquito-borne viral disease whose pathogen is mainly transmitted by females mosquitoes of the species *Aedes aegypti*, the same mosquito that carries the viruses of Zika, yellow fever and chikungunya. As a severe flu-like illness, it may cause

high fever, strong headache, pain behind the eyes, muscles and joints, nausea, vomiting, swollen glands and rash, which can last for one week. There is no specific treatment for dengue fever nor widespread available vaccines and, therefore, control of the disease is mainly performed by suppressing the vector population, as well as identifying outbreaks as quickly as possible [19]. Although case fatality rate can be as low as 1% with proper treatment, the disease comes with great economical and social burden [14].

According to Samir et al. [14], dengue fever is ubiquitous throughout the tropics. The study estimates about 390 million dengue infections worldwide every year, with 96 million of these cases being symptomatic. The Americas contribute with 14% of the symptomatic infections, of which over one fourth occurs in Brazil. This scenario motivates the development of appropriate *early warning systems* (EWSs) to quickly identify new dengue fever outbreaks in Brazilian cities. An EWS is a tool capable of quickly identifying/forecasting risks, and plays a major role on disaster risk reduction by preventing loss of life and mitigating economical and/or material impact [18]. In the context of epidemiology, an epidemiological EWS (EEWS) can be a fundamental step on implementing effective interventions to control infectious diseases, reducing mortality and morbidity in human populations, since it allows authorities to plan ahead and act appropriately.

The main requirement of an EEWS is the availability of up-to-date epidemiological data, which will be used in pattern extraction for forecasting of future incidence of the disease under study. However, epidemiological data cannot be assumed to be provided in real-time. It is common for this kind of data to require time-consuming tests for confirmation of suspected cases. Besides that, the propagation of information from local health care agents to national health authorities may also require extra time. Therefore, it is more reasonable to assume that, at a given moment in time t , the most recent epidemiological data available is associated with time $t - \gamma$, where γ is a positive number. This delay in the report of epidemiological data can be highly detrimental to EEWS, specially for infectious diseases, where current incidence is a valuable information for predicting near future incidence [1, 4, 5].

On the other hand, nowadays a large number of people use text-based online social networks, such as Twitter or Facebook, to discuss various subjects, including personal matters. Similarly, they use search engines and Wikipedia, looking for information. These platforms are capable of providing samples of users' interactions in real-time, leading to the consolidation of useful data sources that may be exploited in an easy and systematic manner. These online data sources have attracted the attention of many researchers who have gone on to use the data to predict real-life events [3, 6, 8, 13,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DH '17, July 02-05, 2017, London, United Kingdom

© 2017 Association for Computing Machinery.

ACM ISBN 978-1-4503-5249-9/17/07...\$15.00

<https://doi.org/http://dx.doi.org/10.1145/3079452.3079464>

15], effectively using people's interaction with these platforms as sensors.

In this work, we combine real-time online data sources with epidemiological data to deal with this time lag required for epidemiological dengue data to be ready for use. We propose a framework that exploits data coming from Twitter to estimate delayed epidemiological dengue data, which is then used together with actual epidemiological data in an EEWS. In our experimental analysis, we compared the proposed framework with alternative approaches and verified that the proposed framework leads to more accurate predictions and that it can be safely applied even on cities for which Twitter data is not abundant, as it automatically identifies scenarios where online data can be usefully employed to improve accuracy.

2 RELATED WORK

The real-time nature of online data makes it attractive for surveillance tasks, particularly in epidemiological contexts, where many strategies that explore this kind of data to estimate incidence of various diseases around the world have been developed. For instance, in [2], the authors used the number of searches on Google containing a specific set of keywords to estimate dengue fever incidence in Singapore and Bangkok, Thailand. Similarly, the authors of [15] developed a methodology that estimates influenza incidence in American cities based on data provided by Google Flu Trends, which provides estimates of influenza incidence based on search activity on Google.

Twitter is also a source of online data frequently exploited for epidemiological surveillance systems. The authors of [3] developed a classifier capable of accurately identifying tweets indicating influenza infection and showed that the volume of these tweets over time presented similar, highly correlated behavior when compared to epidemiological data. A similar idea was investigated in [8], which developed a model for classifying dengue-related tweets into five classes: personal experience (indicating infection), opinion, jokes, information and governmental campaigns. The volume of dengue-related tweets, especially those containing personal experience, was shown to be highly correlated to epidemiological data from Brazil. This relationship was further studied in [16, 17], culminating in the development of models for estimating dengue fever incidence in Brazilian cities.

These works, however, were developed for surveillance, and are not appropriate for EEWS. Although they are capable of estimating incidence of diseases in real-time (nowcast), they do not forecast *future* incidence. The reason behind it is that online data is typically used as input features to (generalized) linear models. Therefore, in order to estimate future incidence, they would require *future* online data. In other words, the fact that online data is obtained in real-time is not sufficient.

In order to predict future outcomes, alternative ways of using online data in EEWS are required. In this direction, a possible approach was evaluated by Generous et al. [6], where access logs of selected Wikipedia articles were used to estimate a linear model for 7 diseases in 9 countries. In order to be able to estimate future incidence, they evaluated the usage of Wikipedia data in a lagged fashion. That is, epidemiological data for time index t would be associated with Wikipedia data for time index $t - \alpha$, where α is a

positive integer. By doing so, the model would be capable of forecasting incidence with up to α time indices of antecedence. In the former, they evaluate values for α up to 4 weeks. In this article, we propose a distinct approach for designing truly predictive models. We obtain estimates of up-to-date epidemiological data, which can then be used within EEWS for forecasting future incidence.

3 DATA COLLECTION

Epidemiological data. We obtained the number of weekly confirmed dengue cases for 5566 Brazilian municipalities from January 2011 to October 2016, summing up to 286 weeks, provided by the Brazilian Ministry of Health. In order to account for different population sizes, we calculated the dengue incidence rate (DIR) as follows:

$$DIR_{s,t} = cases_{s,t} * \frac{100000}{pop_s} \quad (1)$$

where $DIR_{s,t}$ is the DIR at city s during week t , $cases_{s,t}$ is the number of confirmed cases at city s during week t and pop_s is the population size at city s .

The Brazilian Ministry of Health has defined a system of three incidence levels for dengue fever: high, medium and low. High incidence level at a given area occurs when there is more than 300 dengue cases per 100 thousand inhabitants during one month. Medium incidence, on the other hand, means more than 100 dengue cases and less than 300 dengue cases per 100 thousand inhabitants in a month. Finally, low incidence occurs when there is less than 100 cases per 100 thousand inhabitants at the same month. In this work, however, we deal with weekly dengue data. For that reason, we redefined these incidence levels by dividing the number of required cases by 4. Therefore, a high incidence week at a given area implies in a DIR of at least 75, while medium incidence requires DIR between 25 and 75 and low incidence requires DIR smaller than 25.

In order to reduce the effort of evaluating the proposed framework, we consider only cities with more than 100 thousand inhabitants and that reached at least medium incidence during one week in the period under study, resulting in 213 cities. Although small when compared to the initial set of 5566 cities, these cities account for more than 65% of the number of confirmed cases in Brazil. Therefore, cities where an accurate and reliable EEWS is fundamental for controlling dengue epidemics are included in the set of cities under study.

Twitter data. Twitter messages (tweets) were collected by the authors of [8, 16, 17] through the Twitter API, obtaining geolocalized tweets from January 2011 to October 2016 with at least one of the following keywords: *dengue*, *aedes* and *aegypti*.

We aggregated tweets by city and by week in order to obtain time series containing the number of dengue-related tweets posted during each week and at each city under study.

4 INCORPORATING ONLINE DATA INTO EPIDEMIOLOGICAL MODELS

As previously indicated, in most circumstances it is not reasonable to assume that epidemiological data is provided in real time. For this reason, a prediction issued at time t can only explore data available

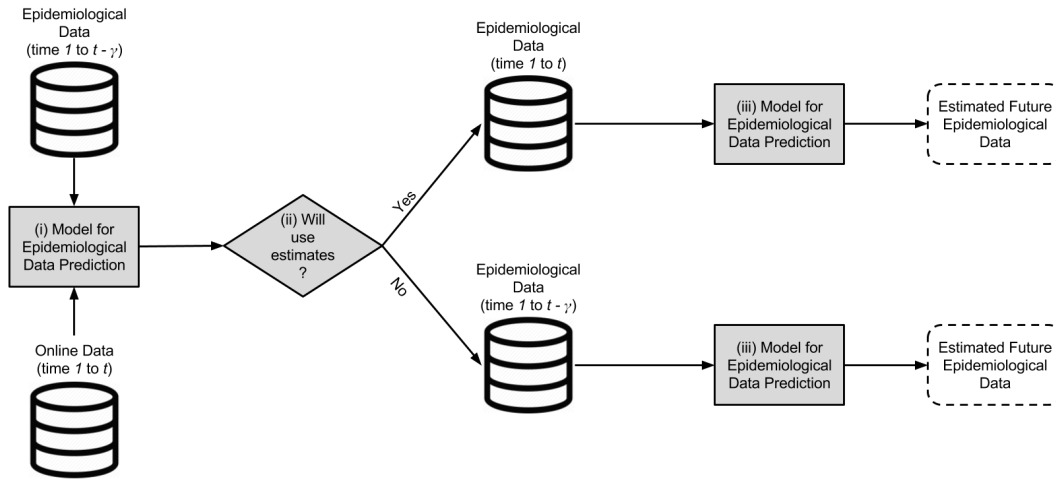


Figure 1: The proposed framework for incorporating Twitter data into an EEWS model.

up to time $t - \gamma$, where γ is the delay associated with the availability of epidemiological data.

Assume that a given EEWS is required to issue predictions with time antecedence of β , i.e., at time t the system would provide predictions associated with time $t + \beta$. Since it would not have access to epidemiological data associated with time between $t - \gamma$ and t , the EEWS may actually use data up to time $t - \gamma$ to forecast incidence at time $t + \beta$, effectively increasing the antecedence for which predictions are issued from β to $\beta + \gamma$. We call this approach the *increased antecedence approach*. The major disadvantage of this approach is that, the larger the antecedence, the more difficult it is to produce accurate forecasts, as incidence of infectious diseases during time t tend to be strongly correlated with incidence in the recent past of t .

In this section, we describe an alternative to deal with delayed epidemiological data. We first provide few definitions key to understanding our proposal:

- The current time is denoted by t , while epidemiological data delay is denoted by γ and predictive antecedence by β .
- We call *delayed epidemiological data* any epidemiological data associated with the time interval $t - \gamma$ and t .
- We call *future epidemiological data* any epidemiological data associated with time $t' > t$.

The proposed approach for incorporating online data into EEWS is described in Figure 1. First, it uses a known relationship between online and epidemiological data (such as a linear or polynomial association) to estimate *delayed epidemiological data* up to time t . Then, it evaluates how reliable are such estimates. If estimates are considered reliable, they are incorporated into available training set. Otherwise, they are simply ignored. Based on the potentially augmented training set, a model is estimated and used to forecast incidence at time $t + \beta$.

As indicated in Figure 1, the proposed framework is composed of three components. The first component is a predictive model

that takes online data as a covariate (e.g., the number of dengue-related tweets) and epidemiological data as the response variable, and outputs estimates of delayed epidemiological data. The second component decides whether estimates of epidemiological data are safe to use, that is, whether the uncertainty associated to estimates is low. Uncertain estimates may introduce extra noise to the training set of the EEWS model, which can be detrimental. The simplest way to define whether estimated data is reliable is to use an uncertainty estimation threshold. Finally, the third component is a traditional EEWS model that takes as covariates data known to be associated with the disease (e.g., temperature, rainfall, time of the year) and epidemiological data as the response variable, and outputs future epidemiological data.

Figure 2 illustrates the use of the framework. Given two time series, one corresponding to the epidemiological and the other to online data, note that the online data is available up to time t , while epidemiological data stops at $t - \gamma$. From these two series, we estimate the epidemiological data from $t - \gamma$ to t . Depending on the confidence of the estimated data, it can be incorporated to the training data of our spatial-temporal model or simply ignored.

The following sections describe how each component of the framework is defined to generate an accurate framework for dengue fever incidence forecasting.

4.1 EEWS for Dengue Fever Incidence Forecasting

We begin by defining the final component of the proposed framework: an epidemiological model to forecast future incidence based on the available training set, possibly extended by integrating estimated delayed DIR values.

Here, we define a model based on [1], where the authors proposed to model log-transformed DIR values of Brazilian cities using a GP model equipped with a covariance function capable of exploiting temporal local dependences typical of infectious diseases, as well as the annual seasonality commonly observed in dengue data [7,

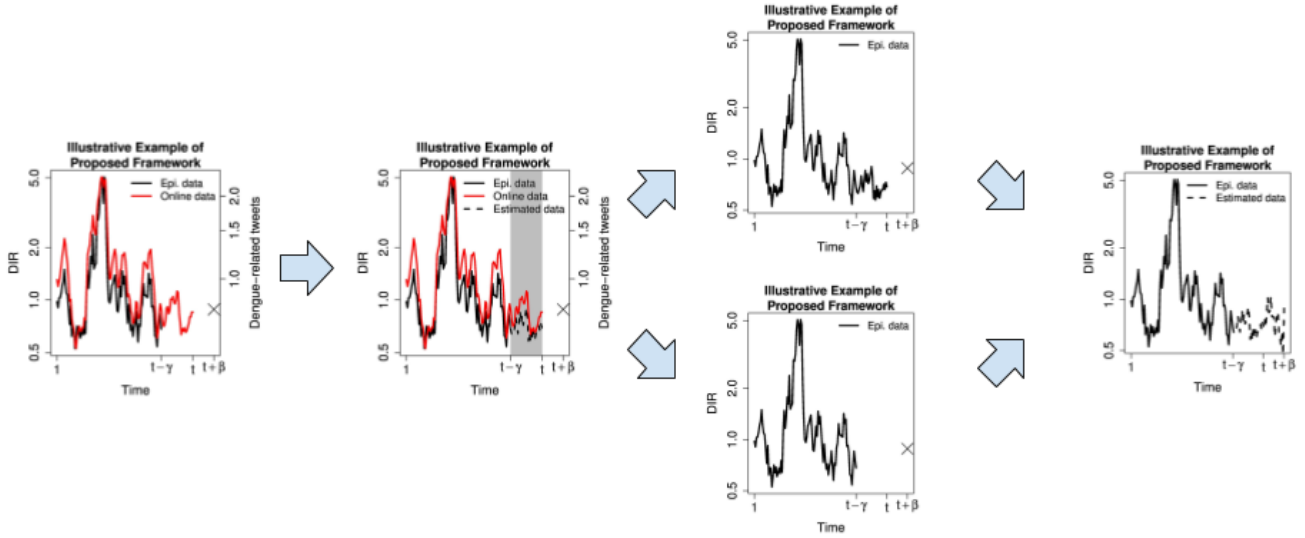


Figure 2: An illustrative example of the use of the proposed framework. At the leftmost figure, we have actual DIR values up to time $t - \gamma$ and the number of dengue-related tweets up to time t (solid black and red lines, respectively). Then, on the next figure, we estimate DIR values from time $t - \gamma$ to t (dashed black line). Following, we have two possible scenarios: estimated DIR values are either incorporated into training data (top) or ignored (bottom). Finally, on the rightmost figure, estimates for future DIR values are made using the available training data.

10, 11]. In summary, we propose to use the following two-part covariance function

$$\begin{aligned}
 k_s(t, t') &= k_{loc}(t - t') + k_{qp}(t - t') \\
 k_{loc}(\tau = t - t') &= \sigma_{loc}^2 \left(1 + \frac{\sqrt{5}\tau}{\ell_{loc}} + \frac{5\tau^2}{\ell_{loc}^2} \right) \exp\left(-\frac{\sqrt{5}\tau}{\ell_{loc}}\right) \\
 k_{qp}(\tau = t - t') &= \sigma_{qp}^2 \left(1 + \frac{\sqrt{5}\tau}{\ell_{qp}} + \frac{5\tau^2}{\ell_{qp}^2} \right) \exp\left(-\frac{\sqrt{5}\tau}{\ell_{qp}}\right) \\
 &\quad \exp\left(-2 * \frac{\sin^2(\pi|\tau|/p)}{\ell_{per}}\right)
 \end{aligned} \quad (2)$$

where $k_s(t, t')$ denotes the covariance between incidence values at city s during weeks t and t' and σ_{loc} , ℓ_{loc} , σ_{qp} , ℓ_{qp} , ℓ_{per} and p are hyperparameters learned from likelihood maximization.

The intuition of the covariance function in Equation 2 is the following. The function is formed by two temporal components, which are used to exploit two temporal patterns, namely local dependences and seasonality. Local dependences are exploited by k_{loc} , which correlates weeks nearby and is expressed in terms of a Matérn covariance function, a covariance function commonly used in the GP literature to enforce smoothness by indicating that nearby data points are highly correlated, while distant data points are loosely correlated [12]. Seasonality is exploited using a quasi-periodic function formulated as the product between a Matérn and a periodic covariance function. By using this formulation, the model is capable of using information from incidence of past years, but is forced to give more relevance to more recent years than very distant ones. Figure 3 shows the covariance function obtained after

hyperparameter optimization (Table 1). Note that covariance is high for τ close to 0, with smaller peaks for τ around 1 year (52 weeks).

Table 1: Hyperparameters obtained after likelihood maximization

	1st Quartile	Median	3rd Quartile
σ_{loc}^2	0.039	0.101	0.186
σ_{qp}^2	0.484	1.427	2.311
ℓ_{loc}	1.000	2.000	7.000
ℓ_{qp}	41.000	58.000	99.000
ℓ_{per}	1.000	1.000	2.000
p	54.000	59.000	73.000

Having defined the covariance function, the model used in this work can be expressed as follows:

$$\begin{aligned}
 DIR_{s,t} &= \exp(y_{s,t} + \bar{y}_s) - 1 \\
 y_{s,\cdot} &\sim \mathcal{GP}(0, k_s(t, t'))
 \end{aligned} \quad (3)$$

where $DIR_{s,t}$ denotes the DIR value at city s during week t , \bar{y}_s indicates the mean of log-transformed DIR values at city s and $k_s(t, t')$ is the covariance function indicated in Equation 2.

4.2 Estimating Delayed Epidemiological Data

The first step on the proposed framework is to exploit a known relationship between online and epidemiological data. Based on the study provided by Gomide et al. [8], which identified a linear association between the number of dengue-related tweets and DIR, and on the results obtained by the Gaussian process-based model

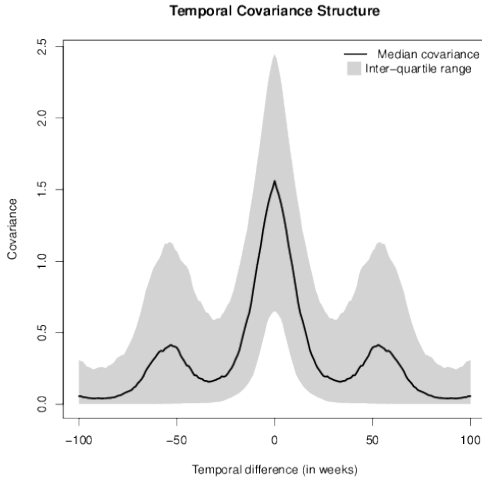


Figure 3: Covariance function considering temporal components in Equation 2 with optimized hyperparameters according to Table 1.

proposed by Albinati et al. [1], we opted for a Gaussian process model equipped with the covariance function indicated in Equation 2, with an additional component designed to exploit the linear association between Twitter and epidemiological data.

Following Albinati et al. [1], we first applied a logarithmic transformation of DIR values to avoid using Gaussian distributions to model count data. We then centered the resulting values and estimated the zero-mean Gaussian process model. In summary, this component is implemented by the following model:

$$\begin{aligned} \text{DIR}_{s,t} &= \exp(y_{s,t} + \bar{y}_s) - 1 \\ y_{s,\cdot} &\sim \mathcal{GP}\left(0, k_s(t, t') + \frac{x_{s,t} * x_{s,t'}}{\ell_{tw}}\right) \end{aligned} \quad (4)$$

where $x_{s,t}$ indicates the log-transformed number of dengue-related tweets posted during week t at city s and $k_s(t, t')$ is given by Equation 2. Note that ℓ_{tw} is an additional hyperparameter, which is also learned from likelihood maximization using available training data.

4.3 Deciding Whether to Use Estimates

In order to avoid using poorly estimated DIR values, which could introduce noise into epidemiological data, we defined an extra module responsible for deciding whether estimated epidemiological data is reliable or not. Since we are assuming a linear relationship between Twitter and dengue data, we opted for using a threshold based on the correlation between epidemiological and Twitter data. Whenever the correlation exceeds the threshold, estimated dengue data is considered. Otherwise, only actual epidemiological dengue data is used. Details on how to define this threshold are given on Section 5.

5 EXPERIMENTAL ANALYSIS

In this section, we show empirical results obtained by the proposed framework. We first define our general experimental setup, and

then assess the accuracy of the proposed framework for incorporating Twitter data into dengue EEWs by verifying if it leads to more accurate predictions than the increased antecedence approach. Then, we provide some discussion on the predictions obtained and on the appropriateness of the proposed framework.

5.1 Experimental Setup

Motivated by the incidence levels defined by the Brazilian Ministry of Health, we evaluate each model according to the area under the receiver operating characteristic curve (AUC) considering incidence levels as labels. The receiver operating characteristic curve is created by plotting the true positive rate against the false positive rate at multiple threshold values, used for defining whether the predicted class is positive or negative. In this sense, it indicates the trade-off between sensitivity and specificity. An area close to 1 indicates that one can increase the model's sensitivity by changing the threshold incurring in few false positive errors, while smaller areas indicate more false positive errors when increasing sensitivity. For each week and city under study, we verify the attained and predicted incidence levels. We calculate the AUC for each level on an one-against-all strategy and then compute the average over the three levels.

Predictions were issued always with $\beta = 4$ weeks of antecedence. Choosing the antecedence value implies in dealing with a trade-off: shorter antecedence leads to more data and potentially more accurate predictions, but less time for health authorities to act in advance to mitigate the impact of future outbreaks. We chose $\beta = 4$ weeks as a balance between utility and accuracy, allowing some time for health authorities to act while still managing to obtain accurate predictions.

We also defined that the first two years (weeks 1 to 104) would be used for training only. Predictions are issued from the third to the sixth year, always with 4 weeks in advance, as indicated previously. This means that our available training set is continuously growing with time, as we collect more data.

Finally, models are always compared using a paired Wilcoxon test with 95% confidence level. Bonferroni corrections were applied whenever multiple statistical tests were required to ensure statistical significance.

5.2 Evaluation of Proposed Framework

In order to use the proposed framework, we must define the correlation threshold that defines whether estimated DIR values are going to be integrated into the training set or not. For that, we instantiated four versions of the proposed framework, each with a distinct threshold value. We compared the difference in accuracy assuming four values of delay in the epidemiological data for each threshold used when compared to simply extending predictive antecedence, as shown in Figure 4. Positive values indicate that the proposed framework achieved better results, while negative values indicate the opposite scenario. Note that, although it is difficult to identify the best threshold value, the proposed framework obtained more accurate predictions for all delays and threshold values used in the experiment, obtaining statistically superior results for delays of at least 4 weeks. Besides that, the difference seems to grow with

larger delays, indicating that the proposed framework is indeed more robust to delays in epidemiological data.

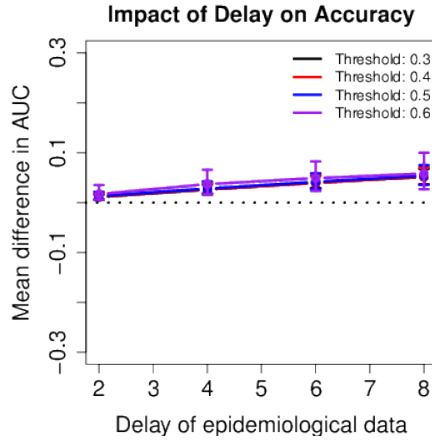


Figure 4: Mean difference in accuracy when comparing the proposed framework with extending predictive antecedence to deal with epidemiological data delay. Whiskers indicate 95% confidence intervals.

Having verified that the proposed framework leads to more accurate predictions in general, we would like to better understand which cities are improved by the proposed methodology and which cities are not. For this analysis, we considered the threshold value of 0.5 and fixed the epidemiological delay as 6 weeks to investigate the role of the amount of Twitter data available on the accuracy improvement obtained. Figure 5 shows that cities that obtained higher improvements in accuracy tend to have more Twitter data available. This result corroborates with the intuitive notion that more Twitter data eases the identification of stronger relationships with epidemiological dengue data.

We now evaluate the spatial distribution of accuracy improvements. Figure 6 shows the spatial distribution of accuracy improvements obtained when applying the proposed framework. Note that improvements are mostly concentrated within the Southeast region of Brazil. Since the Southeast region of Brazil is the most developed region of the country, this distribution reflects the availability of Internet access in Brazil. Having more people capable of posting on Twitter, this region was the most benefited by the proposed framework. We also highlight that, for only 17 cities out of 213, applying the proposed framework was detrimental, indicating that the proposed framework is safe to use regardless of where cities are located.

Although bigger improvements in accuracy were concentrated in the Southeast region of Brazil, we did not observe significant spatial autocorrelation between improvements, as shown in Figure 7.

5.3 Analysis of Predictions

In the last section, we evaluated whether the proposed methodology led to more accurate predictions than the increased antecedence

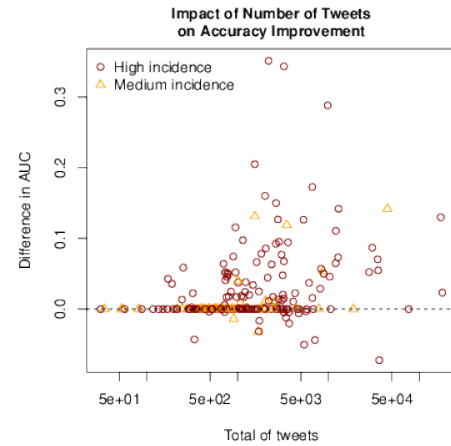


Figure 5: Impact of the amount of Twitter data in the observed difference of accuracy when considering predictions provided by the proposed methodology and provided by the original model with extended predictive antecedence. Each symbol denotes a city, with its shape and color indicating the highest incidence level it achieved in the period under study.

Spatial Distribution of Gains/Losses in Accuracy

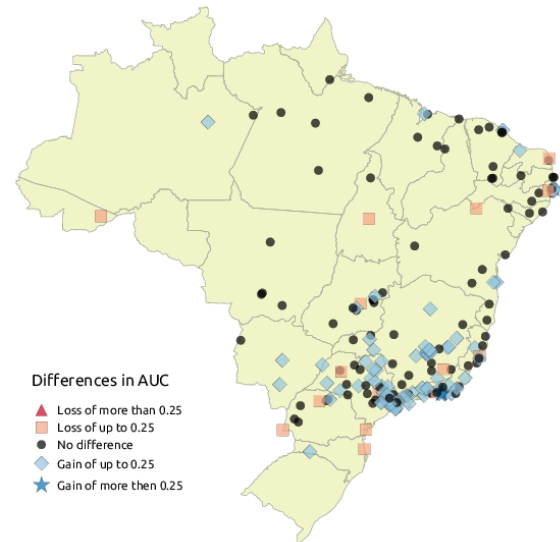


Figure 6: Spatial distribution of the observed difference of accuracy when considering predictions provided by the proposed methodology and provided by the original model with extended predictive antecedence. Each symbol denotes a city, with its shape and color indicating the gain/loss in accuracy provided by the proposed methodology.

approach. We now evaluate the appropriateness of the proposed methodology and analyze the final predictions issued by it.

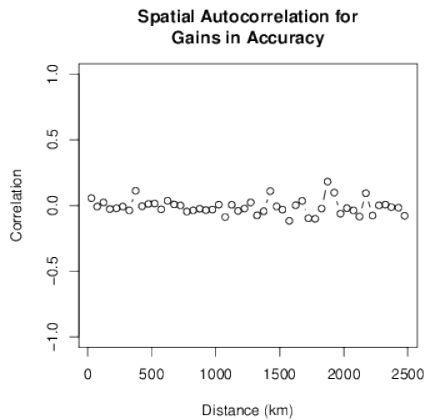


Figure 7: Spatial autocorrelation of the observed difference of accuracy when considering predictions provided by the proposed methodology and provided by the original model with extended predictive antecedence.

We begin by evaluating two assumptions of the proposed model: homoscedasticity and normality of residuals. In other words, the GP-based model assumes that residuals are Gaussian, independent and identically distributed. Figure 8 shows the quantile-quantile plot of residuals. Note that observed quantiles are associated with theoretical quantiles in an approximately linear fashion, indicating that residuals are approximately Gaussian, independent and identically distributed.

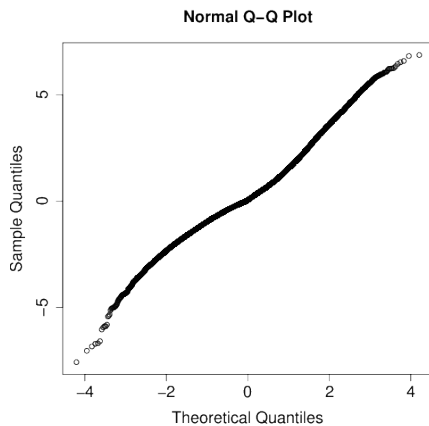


Figure 8: Quantile-quantile normal plot of residuals.

Figure 9 shows empirical cumulative distribution function of AUC. Almost 77% of the cities obtained AUC higher than 0.6, while about 20% of the cities obtained AUC higher than 0.8.

Figure 10, on the other hand, shows the impact of the volume of Twitter data on the accuracy of predictions. Note that the amount of Twitter data is not associated with final predictions. We believe this result is mainly due to the module deciding whether to use

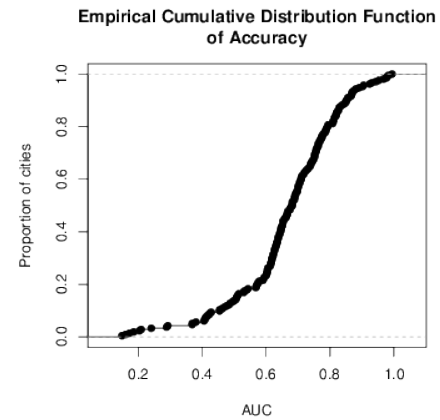


Figure 9: Empirical cumulative distribution function of the observed accuracy when considering predictions provided by the proposed methodology.

estimated DIR values or not, which makes the whole methodology safe to use even when Twitter data is scarce.

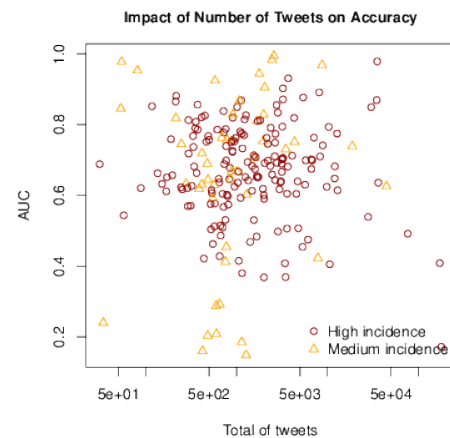


Figure 10: Impact of the amount of Twitter data in the observed accuracy when considering predictions provided by the proposed methodology.

The spatial distribution of final accuracies is shown in Figure 11. Differently from Figure 6, high final accuracies are not strongly concentrated only within the Southeast region. In fact, a significant number of cities from North and Northeast regions obtained high AUC values. These results suggest that, although Twitter was not strongly helpful for cities within these regions, the GP-based EEWS is still capable of issuing accurate predictions with increased predictive antecedence. In this direction, one could argue that the proposed framework is capable of improving predictions exactly where delays in epidemiological data are more detrimental to EEWSs.

Figure 12 shows predictions issued by the proposed framework for all Brazilian capital cities that reached at least medium incidence

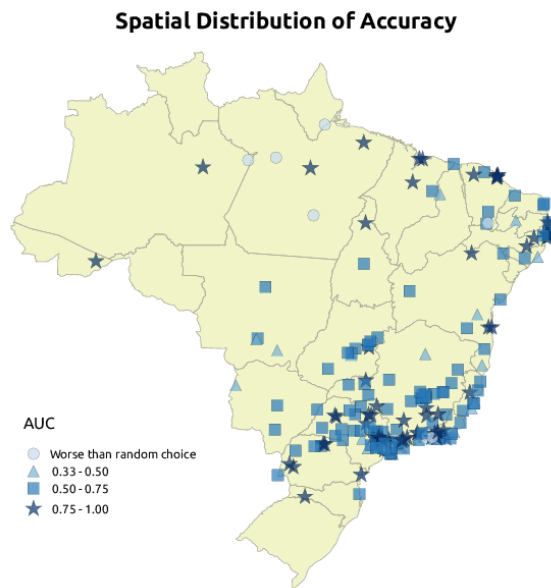


Figure 11: Spatial distribution of the observed accuracy when considering predictions provided by the proposed methodology.

from 2013 to 2016. Cities are sorted by the number of dengue-related tweets collected, where São Paulo was the city with the highest number of tweets and Palmas was the city with the lowest number of tweets. Note that, even for cities with low amount of Twitter data, the proposed framework typically detects major dengue fever outbreaks, reaching the correct incidence level (indicated by the dashed lines). The city of Belo Horizonte, although being the capital city with third highest number of dengue-related tweets, demonstrated to be particularly challenging to forecast, since it exhibited very low incidence during all period under study, except from the beginning of 2013 and 2016, where very drastic outbreaks were observed. However, we note that our proposed methodology, although failing to capture the outbreak in 2013, was able to capture the increase in incidence associated with the outbreak in 2016, reaching the correct incidence level.

6 CONCLUSIONS AND FUTURE WORK

Although EEWS can be employed to forecast DIR, they require up-to-date epidemiological data to work effectively. Unfortunately, epidemiological data requires time to be available, since it may require time-consuming laboratorial tests. Therefore, at any given moment in time t , we cannot expect to have epidemiological data available up to current time. Instead, we would expect to have data available up to time $t - \gamma$. This lack of up-to-date epidemiological data can be highly detrimental to EEWS, as data associated to a recent past is typically very informative, specially for infectious diseases.

In order to mitigate this issue, this work proposes a framework that exploits known relationships between epidemiological and online data, such as data coming from Twitter or Facebook, to

estimate current epidemiological data. Effective EEWS can then be built based on these estimates, without the need of increasing predictive antecedence.

We show that the proposed framework leads to more accurate predictions than employing an EEWS with larger predictive antecedence to account for the delay of epidemiological data. In fact, for delays greater than or equal to 4 weeks, the proposed framework provided statistically more accurate predictions in general. We also show that, although larger amounts of online data led to larger improvements in accuracy, the framework is robust in the sense of being capable of identifying most of the cities for which it was not worthwhile to use Twitter data. For these cities, estimating delayed epidemiological data would introduce noise to the training dataset, thus being detrimental to accuracy. In the end, the proposed framework was able to obtain accurate predictions, with majority of cities obtaining an area under the ROC curve higher than 0.68, when considering that epidemiological data is delayed by 6 weeks.

Our proposed framework obtained higher improvements in accuracy for cities within the Southeast region. According to a census conducted by the Brazilian state-run agency IBGE¹, in 2014, the Southeast region was the region of more houses with Internet access (62.9%), while about only 55% of the Brazilian homes have access to Internet [9]. With more access to the Internet, people within this region are able to produce more dengue-related tweets, which we show to be related to the improvement obtained by the proposed framework. Our experiments, however, suggested that this region is the most affected by the lack of up-to-date epidemiological data. Besides that, it concentrates the majority of dengue cases in Brazil, therefore being one of the Brazilian regions where the proposed framework would be most helpful.

Although improving accuracy in general, the proposed framework has its limitations. First, some cities were not improved due to a weak relationship between online data and epidemiological data. We believe this fact is due to two major reasons: (i) data scarcity and (ii) non-linear associations between dengue and Twitter data. To deal with the former issue, as future work, we intend to aggregate other online data sources, such as data from Google Trends and Wikipedia access logs, which are already being used to estimate current epidemiological data of some diseases. For the latter, we aim to evaluate other models to estimate delayed epidemiological data.

It is important to stress that, although we obtained good predictive accuracy for a large set of Brazilian cities, decision making by public health authorities should not be solely based on estimates provided by the proposed framework, which can be affected by fluctuations on Twitter, among other effects. In fact, our main objective was to develop a system capable of identifying potential future outbreaks, focusing the attention of specialists to specific regions of the country. By doing so, we hope to facilitate and improve their work, since they would be able to better investigate a smaller subset of areas, instead of tracking down DIR around the whole country.

Finally, we would like to highlight that we believe that the proposed framework, although evaluated in the scenario of dengue

¹Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, in Portuguese) – <http://www.ibge.gov.br>

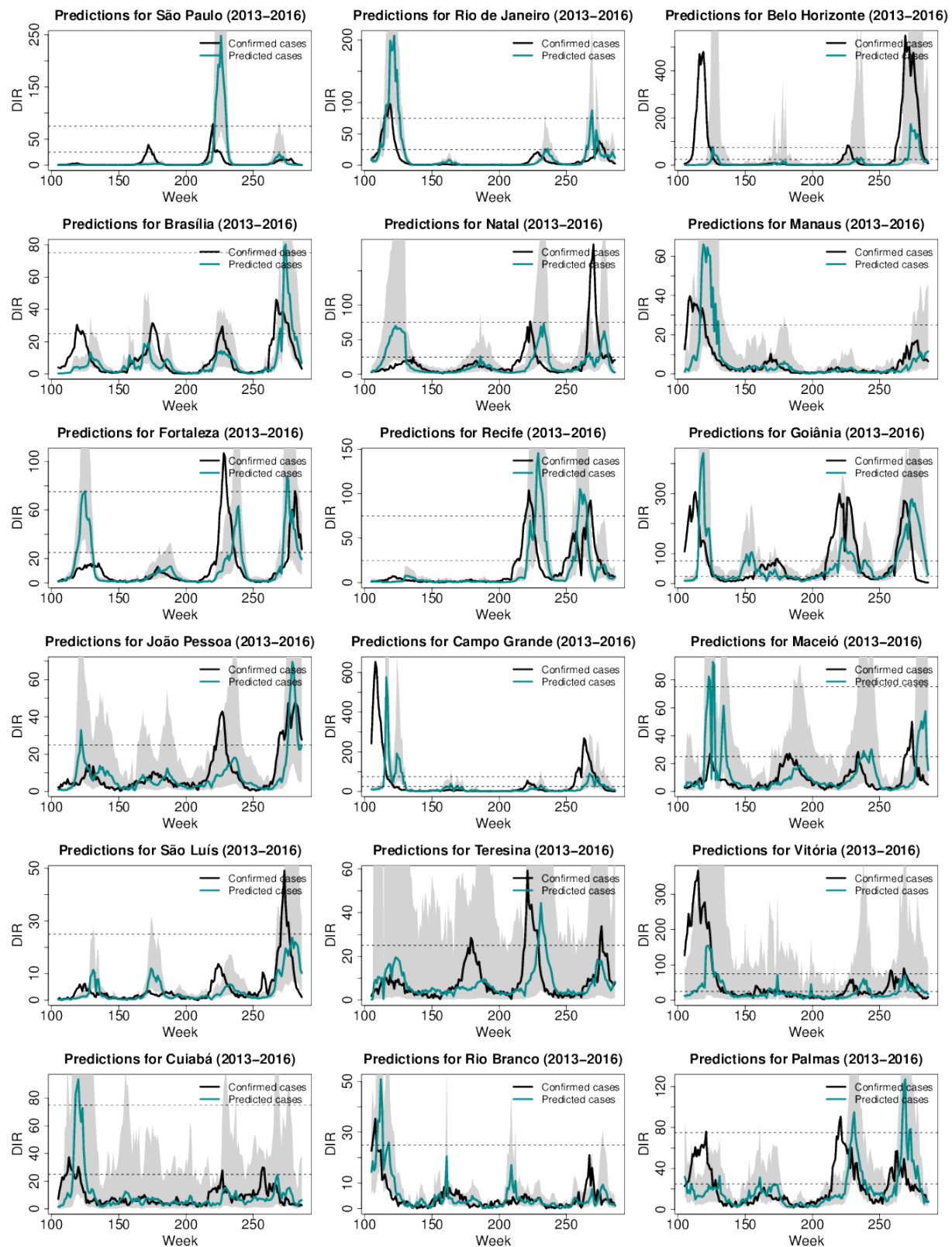


Figure 12: Predictions issued for all Brazilian capital cities under study. The solid black line denotes the observed DIR, while the cyan line indicates the predicted DIR. The gray shaded area exhibits the 95% confidence interval. Dashed lines indicate the boundaries between incidence levels.

fever in Brazilian cities, is flexible and may be applied to other scenarios without large modifications. For addressing this, as future work, we intend to apply it to other diseases, specially mosquito-borne diseases, which typically exhibit similar temporal patterns to dengue fever.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgements

We would like also to thank the authors from Souza et al. [16] for sharing their dataset of geolocalized dengue-related tweets, enabling this study.

This work was partially supported by CNPq, FAPEMIG, CAPES, InWeb, MASWeb, BIGSEA, INCT-Cyber and INCT-Dengue.

REFERENCES

- [1] Julio Albinati, Wagner Meira Jr, and Gisele Lobo Pappa. 2016. An Accurate Gaussian Process-Based Early Warning System for Dengue Fever. In *2016 Brazilian Conference on Intelligent Systems, BRACIS 2016, Recife, Brazil, October 10-13, 2016*.
- [2] Benjamin M. Althouse, Yih Ying Ng, and Derek A. T. Cummings. 2011. Prediction of Dengue Incidence Using Search Query Surveillance. *PLoS Negl Trop Dis* 5, 8 (08 2011), 1–7. <https://doi.org/10.1371/journal.pntd.0001258>
- [3] David A. Broniatowski, Michael J. Paul, and Mark Dredze. 2013. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE* 8, 12 (12 2013). <https://doi.org/10.1371/journal.pone.0083672>
- [4] Nazri Che Dom, A Abu Hassan, Z Abd Latif, and Rodziah Ismail. 2013. Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia. *Asian Pacific Journal of Tropical Disease* 3, 5 (2013), 352–361. [https://doi.org/10.1016/S2222-1808\(13\)60084-5](https://doi.org/10.1016/S2222-1808(13)60084-5)
- [5] Matthew D Eastin, Eric Delmelle, Irene Casas, Joshua Wexler, and Cameron Self. 2014. Intra-and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in Colombia. *The American journal of tropical medicine and hygiene* 91, 3 (2014), 598–610.
- [6] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. 2014. Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Comput Biol* 10, 11 (11 2014), 1–16. <https://doi.org/10.1371/journal.pcbi.1003892>
- [7] Myriam Gharbi, Philippe Quenel, Joël Gustave, Sylvie Cassadou, Guy La Ruche, Laurent Girdary, and Laurence Marrama. 2011. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC infectious diseases* 11 (2011).
- [8] Janaina Gomide, Adriano Veloso, Wagner Meira, Jr., Virgilio Almeida, Fabricio Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. In *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. ACM, New York, NY, USA, Article 3, 8 pages. <https://doi.org/10.1145/2527031.2527049>
- [9] Instituto Brasileiro de Geografia e Estatística. 2014. Pesquisa Nacional por Amostra de Domicílios – PNAD 2014. http://ibge.gov.br/home/estatistica/populacao/acessoainternet2014/default_xls.shtm (2014). [Online; accessed 18-April-2017; in Portuguese].
- [10] Edson Zangiacomi Martinez and Elisângela Aparecida Soares da Silva. 2011. Predicting the number of cases of dengue infection in Ribeirão Preto, São Paulo State, Brazil, using a SARIMA model. *Cadernos de saúde pública* 27, 9 (2011), 1809–1818.
- [11] S Promprou, M Jaroensutasinee, and K Jaroensutasinee. 2006. Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models. *Dengue Bulletin* 30 (2006), 99.
- [12] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. The MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142.
- [13] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 851–860. <https://doi.org/10.1145/1772690.1772777>
- [14] Bhatt Samir, Peter W. Gething, Oliver J. Brady, Jane P. Messina, Andrew W. Farlow, Catherine L. Moyes, John M. Drake, John S. Brownstein, Anne G. Hoen, Osman Sankoh, Monica F. Myers, Dylan B. George, Thomas Jaenisch, G. R. William Wint, Cameron P. Simmons, Thomas W. Scott, Jeremy J. Farrar, and Simon I. Hay. 2013. The global distribution and burden of dengue. *Nature* 496, 7446 (apr 2013), 504–507. <http://www.nature.com/nature/journal/v496/n7446/abs/nature12060.html#supplementary-information>
- [15] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. 2013. Real-time influenza forecasts during the 2012–2013 season. *Nature communications* 4 (2013).
- [16] Roberto CSNP Souza, Denise EF de Brito, Renato M Assunção, and Wagner Meira Jr. 2015. A latent shared-component generative model for real-time disease surveillance using Twitter data. *arXiv preprint arXiv:1510.05981* (2015).
- [17] Roberto C. S. N. P. Souza, Denise E. F. de Brito, Rodrigo L. Cardoso, Derick M. de Oliveira, Wagner Meira Jr., and Gisele L. Pappa. 2014. An Evolutionary Methodology for Handling Data Scarcity and Noise in Monitoring Real Events from Social Media Data. In *Advances in Artificial Intelligence - IBERAMIA 2014 - 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24-27, 2014, Proceedings (Lecture Notes in Computer Science)*, Ana L. C. Bazzan and Karim Pichara (Eds.), Vol. 8864. Springer, 295–306. https://doi.org/10.1007/978-3-319-12027-0_24
- [18] Alison Wiltshire. 2006. “Developing early warning systems: A checklist. In *Proc. 3rd Int. Conf. Early Warning (EWC)*.
- [19] World Health Organization Media Centre. 2016. Dengue and severe dengue. <http://www.who.int/mediacentre/factsheets/fs117/en/>. (2016). [Online; accessed 25-April-2016].