Predicting Dengue Incidence Leveraging Internet-Based Data Sources. A Case Study in 20 cities in Brazil.

Gal Koplewitz[1,2], Fred Lu[3,4], Leonardo Clemente[2], Caroline Buckee[2], Mauricio Santillana[3,6,#]

[1] Harvard J. A. Paulson School of Engineering and Applied Sciences, Cambridge, MA
[2] Harvard T. H. Chan School of Public Health, Boston, MA
[3] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA
[4]Department of Statistics, Stanford University
[6] Department of Pediatrics, Harvard Medical School, Boston, MA
[#] Correspondence: msantill@g.harvard.edu

Abstract

The dengue virus affects millions of people every year worldwide, causing large epidemic outbreaks that disrupt people's lives and severely strain healthcare systems. In the absence of a reliable vaccine against it or an effective treatment to manage the illness in humans, most efforts to combat dengue infections have focused on preventing its vectors, mainly the Aedes aegypti mosquito, from flourishing across the world. These mosquito-control strategies need reliable disease activity surveillance systems to be deployed. Despite significant efforts to estimate dengue incidence using a variety of data sources and methods, little work has been done to understand the relative contribution of the different data sources to improved prediction. Additionally, most work has focused on prediction systems at the national level, rather than at finer spatial resolutions. We develop a methodological framework to assess and compare dengue incidence estimates at the city level and evaluate the performance of a collection of models on 20 different cities in Brazil. The data sources we use towards this end are weekly incidence counts from prior years (seasonal autoregressive terms), weekly-aggregated weather variables, and real-time internet search data. We find that a random forest-based model effectively leverages these multiple data sources and provides robust predictions, while retaining interpretability. For real-time predictions that assume long delays (6-8 weeks) in the availability of epidemiological data, we find that real-time internet search data are the strongest predictors of Dengue incidence, whereas for predictions that assume very short delays (1-2 weeks), short-term and seasonal autocorrelation are dominant as predictors. Despite the difficulties inherent to city-level prediction, our framework achieves meaningful and actionable estimates across cities with different characteristics.

Author Summary

As the incidence of infectious diseases like dengue continues to increase throughout world, tracking their spread in real time poses a significant challenge to local and national health authorities. Accurate incidence data are often impossible to obtain as outbreaks emerge and unfold, and a range of nowcasting tools have been developed to estimate disease trends using different mathematical methodologies to fill the temporal data gap. Over the past several years, researchers have investigated how to best incorporate internet search data into predictive models, since these can be obtained in real-time. Still, most such models have been regression-based, and have tended to underperform in cases when epidemiological data are only available after long reporting delays. Moreover, in tropical countries, these models have previously been tested and applied primarily at the national level. Here, we develop a machine learning model based on a random forest approach and apply it in 20 cities in Brazil. We find that our methodology produces meaningful and actionable disease estimates at the

city level, and that it is more robust to delays in the availability of epidemiological data than regression-based models.

Introduction

Dengue fever is one of the fastest-growing mosquito-borne viral diseases in the world. With an estimated 390 million infections each year, dengue threatens roughly 3.9 billion people in 128 countries and poses a growing health and economic problem throughout the tropical and sub-tropical world.[1] As climate change and urbanization intensify, the geographic range of dengue is expected to spread even further.[2] Though the disease often manifests asymptomatically, severe cases can lead to hemorrhage, shock and death.[3]

Health services have thus strained to address the burden of dengue morbidity and mortality through a variety of means. Without a reliable vaccine or an effective treatment to manage the illness in humans, one effort, promoted by the World Health Organization (WHO), has aimed to achieve better early case detection. By focusing on improving epidemiological surveillance and attaining more timely identification of outbreaks, public health officials hope that preventive measures to reduce the spread of the disease can be used more effectively (vector control methods include, for example, the distribution of mosquito nets). However, effective real-time tracking of the spread of dengue – let alone prediction – has proven difficult. This is particularly evident in sprawling countries like Brazil, in which health resources are spread thin over a vast range of localities in which dengue is endemic. Governments typically rely on clinic-based reporting for case counts, but in Brazil (as in other countries) this information is often lagged in time and subject to post-hoc revisions, thus limiting the potential effectiveness of interventions.[4,5] Thus, the development of data-informed tools for dengue surveillance which provide accurate case counts in real-time has increasingly become a priority.

The transmission dynamics of dengue and the time scales at which they occur lend themselves to tracking patterns of infection. In tropical environments, *Aedes aegypti* and *Ae. albopictus* mosquitoes can transmit dengue viruses within a week of infection. Once infected by a mosquito, a person can become ill within a week, and show symptoms for up to 10 days (other mosquitos can subsequently pick up dengue from an infected person within a 5-day window).[6, 7] A range of external conditions have also been shown to affect dengue transmission. Among these are precipitation, temperature and other seasonal weather patterns, which influence the spread of the disease by affecting the development and lifespan of the dengue-carrying mosquitos.[8, 9,10, 11, 12] Additional factors include the human population density in a given town or region, as well as the degree to which various mosquito control efforts have been implemented by local health authorities.[13, 14]

Harnessing these various factors, a large number of models have been developed over the years in the attempt to forecast or nowcast dengue incidence. These range from compartmental mechanistic models, based on a set of differential equations, to statistical autoregressive models such as Seasonal Autoregressive Integrated Moving Average (SARIMA), which leverage both seasonal patterns and recent trends to produce disease estimates.[15, 16, 17, 18, 19] Over the past few years, search activity on internet search engines has increasingly been explored as a potential data source for these models. As internet access in the developing world increased, researchers have shown the potential of applying user activity data from search engines and social media to make predictive estimates of dengue incidence levels.[29,33]

However, much of the work in this field has been done at a national level, with models estimating disease levels over vast geographical swaths with highly varying local conditions and rates of disease. This variability, in addition to smaller population sizes and fewer reliable data sources, makes modeling disease rates at the municipal level more technically challenging. But in order to be useful to local and national health administrators (as well as to international health organizations), accurately predicting incidence at the city-level is crucial—for example, in guiding the distribution of resources such as mosquito nets.

In recent years, more attempts have been made to fill this gap. In Brazil, a joint effort by academics and health officials has produced "Infodengue," a system for dengue surveillance at the city level which has been running since 2015.[20] Delays and inaccuracies in reported disease surveillance data are some of the key difficulties in the tracking of epidemics, and a number of approaches, such as Bayesian hierarchical modelling and constrained P-spline smoothing, have been used by researchers in the attempt to account for these delays and the uncertainty they introduce.[21, 22, 23] Other efforts to mitigate the effect of delays in reporting have sought to incorporate novel real-time data sources, such as Twitter activity, in order to improve nowcasting model performance.[24, 25]

Our contribution. We extend methodological frameworks, previously used for flu surveillance, to estimate dengue activity at the city level up to 8 weeks ahead of the publication of epidemiological reports. We assess the predictive performance of a collection of models by comparing their estimates, produced in a strictly out-of-sample fashion (only using information that would have been available at the time of prediction), with the subsequently observed dengue incidence. The underlying statistical methods we compare are both regression-based (LASSO) and non-parametric ensembles (Random Forest), and the data sources we leverage for these estimates are: (a) weekly incidence counts from prior years (seasonal autoregressive terms), (b) weather measurements, and (c) real-time dengue-related Google Search Trends data. We evaluate the performance in tracking dengue in 20 cities in Brazil and highlight the conditions in which this framework achieves more accurate predictions. Our results show that despite the difficulties inherent to predictions at the city level, our framework achieves meaningful, actionable estimates, and highlights the conditions in which our models perform most accurately. Finally, we find that our approach is capable of identifying whether or not an upcoming season will experience an epidemic with accuracies above 75%, up to 8 weeks ahead of available reports.

Materials and Methods:

Data

We used three distinct sources of information for our study: (a) historical dengue incidence from Brazil's Ministry of Health, (b) Google search frequencies of dengue-related queries, aggregated at the state-level, for the states in which the 20 chosen cities are located, and (c) Weather data, obtained from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2).[26]

We analyzed weekly dengue activity in 20 cities in Brazil: Aracaju, Barra Mansa, Barretos, Barueri, Belo Horizonte, Eunápolis, Guarujá, Ji Paraná, Juazeiro do Norte, Manaus, Maranguape, Parnaiba, Rio de Janeiro, Rondonópolis, Salvador, Santa Cruz do Capibaribe, São Gonçalo, São Luís, São Vicente, Sertãozinho, and Três Lagoas. We chose these Brazilian cities based on several criteria. First, they all had populations over 100,000 by July 2016 (the end of the time range we examined) and varied widely in population size above that threshold. Second, the cities were all chosen to be "dengue endemic" locations, experiencing between 7 and 10 epidemic

years between 2001 and mid-2016 (following the definition of the Brazilian Ministry of Health, an epidemic year is one in which the number of confirmed cases of dengue fever exceeds 100 per 100,000 persons[27]). Finally, they were chosen from a wide geographic range of 13 different states in Brazil and have a wide range of population densities, both of which are epidemiological factors known to influence disease dynamics. For the full summary of the characteristics of the different cities, see Table B in S1 text.

Epidemiological data. Weekly dengue case counts from January 2010 to July 2016 were obtained from the Ministry of Health of Brazil. We confirmed that the ministry-reported annual totals match the sum of case counts over each year.

Online search volume data. Weekly Google search frequencies for dengue-related queries were obtained from Google Trends (www.google.com/trends) using the Google Health Trends API. The Google Trends API was accessed using the gtrends-tools interface (https://github.com/fl16180/gtrends-tools). The search terms were downloaded at the state-level, for the states in which each of the 20 cities is located (Google Trends data at the city-level are not currently available in Brazil).

For online search term selection, we initially sought to use Google Correlate (www. google.com/correlate), which is designed to identify search terms correlating highly with a given time series. This method has been used in the past with success[28]. However, since most of the search terms returned by Google Correlate for our time series of dengue incidence were unrelated to dengue, and since it was discontinued in the course of our work (in December of 2019), we instead used the Google Trends (www.google.com/trends) tool to identify queries which are highly correlated with the term 'dengue' (a feature enabled by the Google Trends interface). In order to ensure the model was robust and generalizable, we ignored terms unrelated to dengue, and verified the terms with a native Portuguese speaker. The weekly aggregated search frequencies of these terms were then downloaded within the time period of interest. Importantly, since we intended the method to generalize to states and cities across Brazil, we used the same terms for the 20 cities. The query terms are presented in Table A in S1 text.

Climate data. Climate data were collected from MERRA-2 (Modern Era Retrospective-analysis for Research and Applications). The MERRA-2 data are publicly available through the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center. For each of the 20 cities, daily climate indicators from Jan 1 2000 to Dec 31 2016 were created, with the following features: 2-meter air temperature (K), precipitation (mm), mean wind speed (m/s), and 2-meter specific humidity (kg/kg). These data were then aggregated into weekly reports, in the range of dates between January 2010 and July 2016, to align with the epidemiological dengue incidence data.

Methods

Our model draws on a range of data sources that have been used in the multivariate linear regression modeling framework ARGO (AutoRegressive model with GOogle search queries as exogenous variables), previously used to track flu incidence using flu-related Google searches[29]. But the underlying machine learning methodology in our model differs fundamentally, and we extend other aspects of previous models significantly. We introduce Random Forest-based prediction in addition to previously tested L1-based (LASSO) regularized regression models. This new model was used to combine information from historical dengue case counts and dengue-related Google search frequencies, as well as climate data, with the goal of estimating dengue activity at different time ranges ahead of the publication of official health reports.

At a high level, our models are re-trained each week on data available at the time of prediction in order to estimate an out-of-sample nowcast of dengue incidence for that week. The weekly generated training sets consisted of a growing time-window which contained incidence data from time points up to 8, 6, 3 or 1 weeks prior to the time of estimation. The minimal window size contained 52 time points (a full year), and the maximal contained over 300, when estimating some of the final points in our range (in mid-2016). This growing window approach allowed the model to constantly improve its predictive ability by taking into account an ever-larger sample of the relationship between internet search behavior, weather, and dengue activity. An alternative approach, using a moving window of a constant size, proved to perform less well in most cases. For completeness in our modeling approaches, we also incorporated information on dengue activity from one, two and three years before the time-to-prediction, to test if long-term seasonal activity would improve performance as the literature has suggested.[15, 29]

Model formulation and assessment. Our models were based on the assumption that when there are more dengue cases, more dengue-related searches will be observed. This is formalized mathematically via a hidden Markov model, as explained in Yang et al, 2015.[29]

Assuming that epidemiological reports were available with different time delays ranging from 1 to 8 weeks, we constructed models that would only have access to the most recent information available at the time of prediction. Thus, our models incorporated historical information (in the form of autoregressive features) from the prior 52 weeks, if available, or from a reduced set depending on the assumed delay in the availability of epidemiological information. In other words, taking J to be the number of weeks for which we incorporate incidence data as autoregressive features, we defined four different set-ups: $J_1 = \{8, 9, \ldots, 52\}$, $J_2 = \{6, 7, \ldots, 52\}$, $J_3 = \{3, 4, \ldots, 52\}$, $J_4 = \{1, 2, \ldots, 52\}$. For J_1 epidemiological reports are available with an 8-week delay, for J_2 with a 6-week delay and so on. These choices of J capture the influence of short-term fluctuations, which has been shown to be strongly predictive for dengue case counts.[29] The effect of long-term seasonality is also considered, implicitly and explicitly, by the inclusion of our expanding training window strategy, which incorporates new training samples as more data is collected every week, and by explicitly including as predictors weeks 78, 104, and 156 whenever they were available. Finally, we define K as the set of Google query terms collected from Google Trends, and $X_{k,t}$ as the dengue-related Internet time series of the k^{th} term in K at a time t.

Model parameter estimation.

LASSO Regression. The Least Absolute Shrinkage and Selection Operator (LASSO) is a linear regression technique that minimizes the residual sum of squares subject to a L1 norm.

At a given time t, we estimate the log-transformed case counts y_t, $y_t = \log(c_t + 1)$, to be

$$y_t = \mu_y + \sum_{j \in J} \alpha_j y_{t-j} + \sum_{k \in K} \beta_k X_{k,t} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

where α_j and β_k are the estimation coefficients for y_{t-j} and $X_{k,t}$ correspondingly, μ_y is an intercept term and ϵ_t is an error term. The L1 norm is a regularization technique that imposes a constraint over α_j and β_k, making the

sum of the absolute value of the linear coefficients to not exceed a specific value (this value is a hyper-parameter, and is found via 5-fold cross validation).

As a linear model, the coefficients associated with each feature are highly interpretable. L1 regularization also performs feature selection, zeroing out coefficients of features that contribute little to the predictions for each time window.

Random Forests

Random Forests are a classification and regression method based on decision trees, models in which the final outcome – either a class or continuous variable – is determined after a sequence of comparisons of the values of predictors against threshold values. Each comparison and subsequent branching in a decision tree corresponds to a partition of the feature space with a line, plane or hyperplane. Learning a decision tree model means producing an optimal partition of the feature space, which produces the "purest" regions and minimizes classification or regression error (learning the smallest such optimal tree is an NP-complete problem, but "reasonably optimal" models can be constructed using a greedy algorithm). Regression trees, once constructed, define simple step functions, which are constant on the partitions of the feature space. If sufficiently complex, such partitions can approximate complex non-linear functions.

However, large and complex decision trees are prone to overfitting and high variance. This can be amended by using Random Forests, a form of bagging ("bootstrap aggregating") in which multiple trees are trained on random samples of the training data – and then for a given input, the output is the averaged output of those trees.[30, 31] To ensure the ensemble of decision trees is independent, for each split of each tree a random subset of predictors P' is selected from the full set of predictors P. Finally, Random Forests have the advantage of being relatively interpretable, as widely accepted methods exist for calculating the relative importance of predictors in a "trained" forest (see [v], as well as [32]). Still, they are not as intuitively interpretable as simple decision trees or linear models, in which one can more explicitly infer how the response variable changes in response to specific changes in features X.

All statistical analyses were performed with Python, version 3.6.4, in Jupyter notebook, using the statistical and machine learning libraries NumPy, Pandas, and Scikit-Learn.

Benchmark Models and Feature Sets

To our knowledge, few previous attempts were made to forecast or "nowcast" dengue incidence at the city level in Brazil. One such instance, which harnessed data from twitter to make estimates at both the country and city levels, found that tweets were useful for both forecasting and nowcasting dengue cases at the city level, though the association between the two was not as strong as at the country level.[25, 33] Another such study focused on applying time-series analysis comparatively between two particular cities, Recife and Goiania, which have populations of a similar size.[34] The Brazilian health authorities themselves typically release case counts 2-4 weeks after the fact, and frequently correct these figures substantially weeks after the initial publication. Thus, there was no clear external baseline with which to compare our results.

To evaluate performance with different assumptions about the availability of data and the relative contributions of various features, we constructed a number of internal benchmarks. First, we compared four different feature sets from our data sources: one solely with Google Trends data (which we label GT), a second solely with autoregressive data (AR), a third which included both (AR + GT), and a fourth that also took into account the

climate data of each week and the week prior to it (AR + GT + W). In this way, we could assess the impact of each of the data sources at predictions with different models from different time horizons.

Second, we compared our two statistical methodologies, regression-based (LASSO) and non-parametric ensemble (Random Forest), and assessed how they performed relative to one another across the different feature sets and from different time horizons. In particular, we assessed the Random Forest model against the regression methodologies, which have been much better studied in the context of disease incidence nowcasting applications.

More generally, we evaluated which models and which data sources perform best at each time point with each methodology, while also summarizing performance across these, in order to determine which methodology and feature set were most robust, and which led to the strongest performance across the board.

Model assessment

We generated model estimates over the period between January 2011 and July 2016 with all of our models for each of the 20 cities, as selected following the previously described procedure. We used the following metrics to assess the performance of our models: root mean square error (RMSE), relative RMSE (R-RMSE), the R-squared coefficient of determination (R^2) and the Pearson correlation coefficient. These were computed for the entire prediction period, over weekly intervals.

For each model, we also tested four variants based on simulating how recently the last official dengue case count report was received (denoted as 1, 3, 6, and 8-weeks before the "current," predicted dengue report). Since the time delay between official case count reports is variable, it is important to assess how robust the models are to varying availability of autoregressive information.

Finally, to analyze more fully the long-term influence that historical dengue activity has on the future dynamics of outbreaks, we compared our selected AR model with an enhanced AR model, which included additional seasonal autoregressive features characterizing historical dengue activity (occurring up to 3 years in the past). Our results, which can be seen in Figure A and Table C of the **S1 text**, were effective in some cities but not in others, and so were not incorporated into the final model.

Utilizing dengue activity point estimates to predict an incoming epidemic in Brazil

Building on the primary model for nowcasting real-time dengue incidence, we also tested our ability to predict, as a *binary* task, whether or not an epidemic would occur as a dengue season unfolds. More specifically, for each of the 20 municipalities, we assessed whether the cumulative number of dengue cases (a mix of available reported epidemic observations and disease estimates produced by our models) crossed a specified threshold value, referred to as the epidemic threshold, on a weekly basis. As the assumed delay in the availability of "observed" epidemiological information is up to 8 weeks, we substituted the 8 most recent weekly "missing" reports using our dengue point-estimates, and aggregated them along with the current "observed" available information as to increase our ability to predict a potential epidemic every week. Specifically, if the cumulative number of cases for a given time interval t_e exceeded the epidemic threshold value, we labelled the interval as epidemic. If it did not, we labelled it as non-epidemic. If our model using our substituted point estimates successfully predicted an epidemic within a dengue season as defined by the cumulative official case counts, we considered that season as a true positive. If the model did not predict an epidemic during all its weekly assessments and this remained consistent with the official epidemiological data, we considered that case a true

negative. We generated the binary classification dataset by dividing the historical dengue activity time-series of each municipality into 52-week time intervals. These time intervals empirically center the high dengue activity periods, and keep the inter-outbreak activity (seasons with low dengue activity) at the start and the end of each interval. For each time interval, the cumulative dengue activity was calculated: from 0 in the first week, t_0, to the total number of cases at week 52, or t_{52}.

Given that the distribution of epidemic and non-epidemic intervals depends on the selection of the epidemic threshold – we tested and repeated our task using a range of values consistent with the standard thresholds reported in the literature, from 100/100,000 to 300/100,000.

Results

Across almost all of the 20 cities (19/20), with varying population sizes and regions, we found that our models accurately estimated dengue incidence when recent case count information is available. When longer delays in the availability of epidemiological data were assumed, the LASSO-based model slightly outperformed the Random Forest-based models, and the best-performing feature set was GT. This advantage narrowed in scenarios which assumed shorter delays, of 1-3 weeks in advance, in which cases the two underlying methodologies tended to perform comparably. The Random Forest-Based model, however, was more robust to changes in features and assumptions about the availability of real-time epidemiological data. It also tended to produce fewer extreme values (see **Fig 1**).

As assumed delays in the availability of epidemiological data grow smaller, performance improves across the board, with lower RMSE and higher Pearson correlation observed in all models. For predictions that assume very short delays in the availability of epidemiological data, short-term and seasonal autocorrelation were key to improving estimates and captured a substantial amount of dengue variability. For predictions that assume longer delays, the real-time Google search trends data captured the most substantial amount of dengue variability. To highlight these effects, we examine a number of cities, and focus on the model that tended to be most robust across different feature sets: the underlying RF methodology, with AR + GT feature set. In **Fig 2**, we show nowcasts in four cities using this model: Sao Luis, Belo Horizonte, Barra Mansa and Maranguape. These cities were chosen based on their different population sizes, peak epidemic rates, and weather patterns, and so best show the robustness of the model (see **Table B** in S1 text).

We focus on one of these, the city of Barra Mansa in the State of Rio de Janeiro, to highlight performance at a more granular level and allow comparisons between the different metrics, feature sets and availability of epidemiological data (see **Table 1**). Barra Mansa was chosen because its density, area and population size are all close to the median of the 20 cities, and because its performance metrics demonstrate some of the trends observed elsewhere. Data from all 20 cities are available at this resolution in **S2 spreadsheet**.

Figure 1. Performance across cities, Pearson Correlation and Relative RMSE

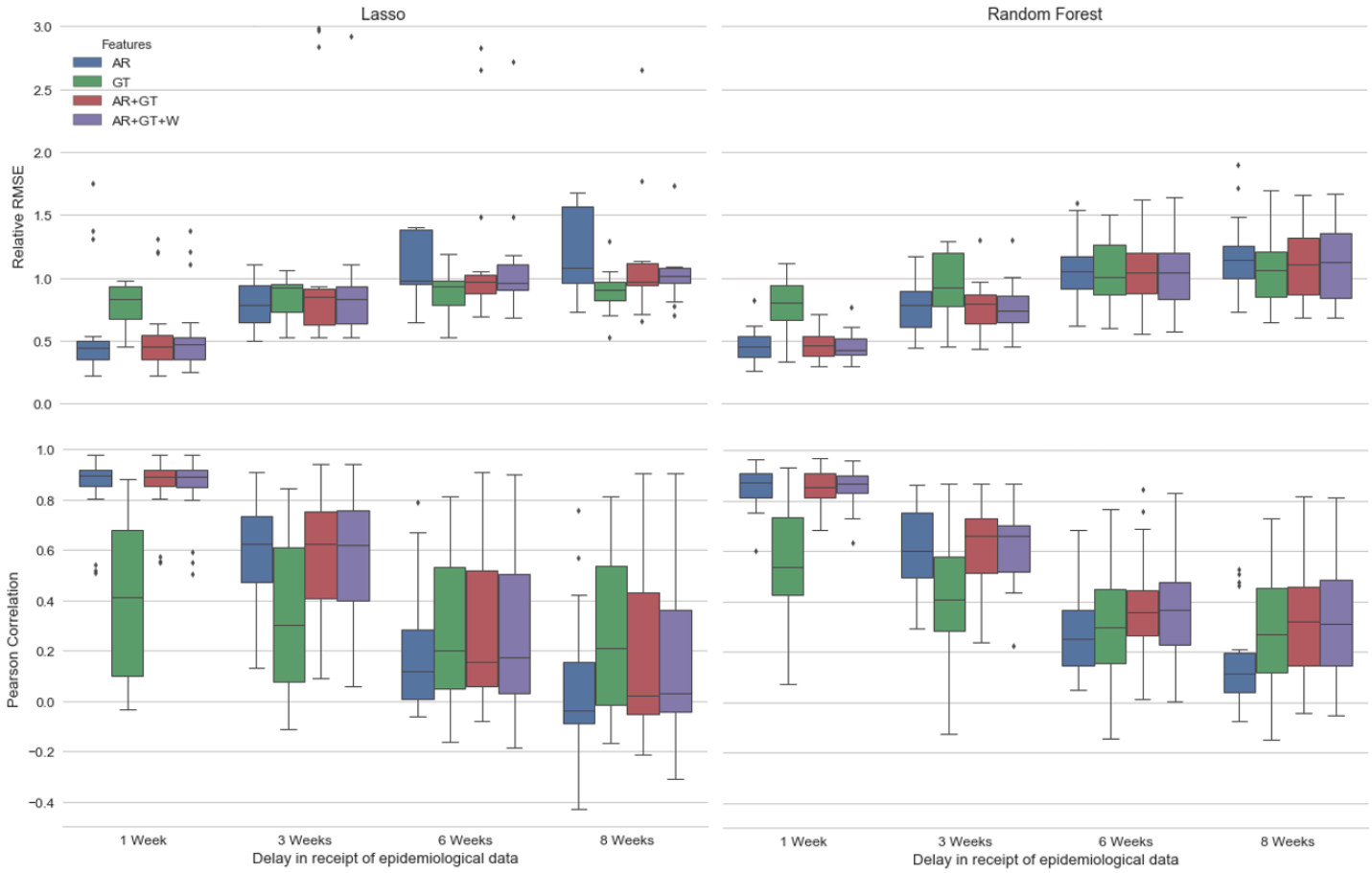


Figure 2. Predictions for 4 cities, different delays in the availability of epidemiological data.

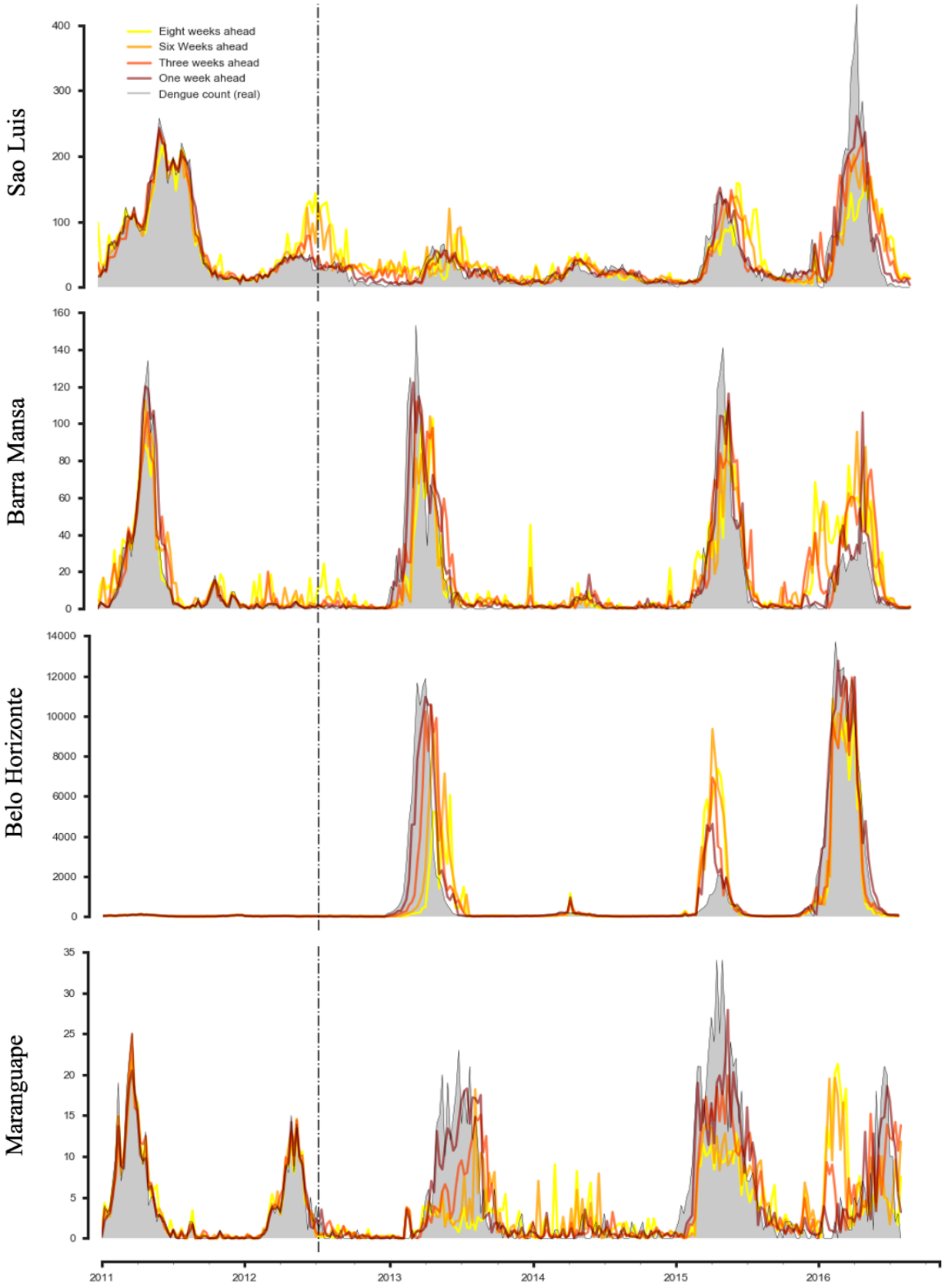


Table 1 Performance of Dengue incidence prediction models from different time horizons, for time period of January 2011 to July 2016, in the city of Barra Mansa, State of Rio de Janeiro, Brazil

Model	AR Lag	Features	RMSE	Relative RMSE	R^2	Pearson Correlation
Lasso Regression	**8 weeks**	AR	28.425	0.897	0.009	0.188
		GT	**23.606**	**0.745**	**0.317**	**0.563**
		AR+GT	23.828	0.752	0.304	0.555
		AR+GT+W	24.67	0.778	0.254	0.505
	6 weeks	AR	26.65	0.845	0.122	0.355
		GT	23.546	0.746	0.315	0.562
		AR+GT	**22.615**	**0.717**	**0.368**	**0.608**
		AR+GT+W	23.039	0.73	0.344	0.587
	3 weeks	AR	19.264	0.615	0.536	0.733
		GT	21.581	0.689	0.418	0.649
		AR+GT	18.859	0.602	0.556	0.752
		AR+GT+W	**18.789**	**0.6**	**0.559**	**0.753**
	1 week	AR	12.485	0.4	0.804	0.897
		GT	20.229	0.649	0.485	0.703
		AR+GT	12.259	0.393	0.811	**0.901**
		AR+GT+W	**12.222**	**0.392**	**0.812**	**0.901**
Random Forest	**8 weeks**	AR	26.382	0.832	0.146	0.508
		GT	**21.061**	**0.664**	**0.456**	**0.68**
		AR+GT	23.355	0.737	0.331	0.596
		AR+GT+W	22.514	0.71	0.378	0.631
	6 weeks	AR	24.625	0.781	0.251	0.591
		GT	22.203	0.704	0.391	0.642
		AR+GT	22.055	0.699	0.399	0.664
		AR+GT+W	**21.154**	**0.671**	**0.447**	**0.678**
	3 weeks	AR	18.332	0.585	0.58	0.776
		GT	21.07	0.673	0.445	0.676
		AR+GT	**17.613**	**0.562**	**0.612**	**0.793**
		AR+GT+W	19.354	0.618	0.532	0.749
	1 week	AR	11.047	**0.354**	0.846	0.92
		GT	19.408	0.622	0.526	0.729
		AR+GT	**11.027**	**0.354**	**0.847**	**0.924**
		AR+GT+W	11.844	0.38	0.823	0.91

Figure 3: Comparison of Random Forest and Lasso model performance. Mean taken across the different cities, with different delays in availability of epidemiological information (AR1 to AR8) and different feature sets (AR, GT AR+GT, AR+GT+W) shown.

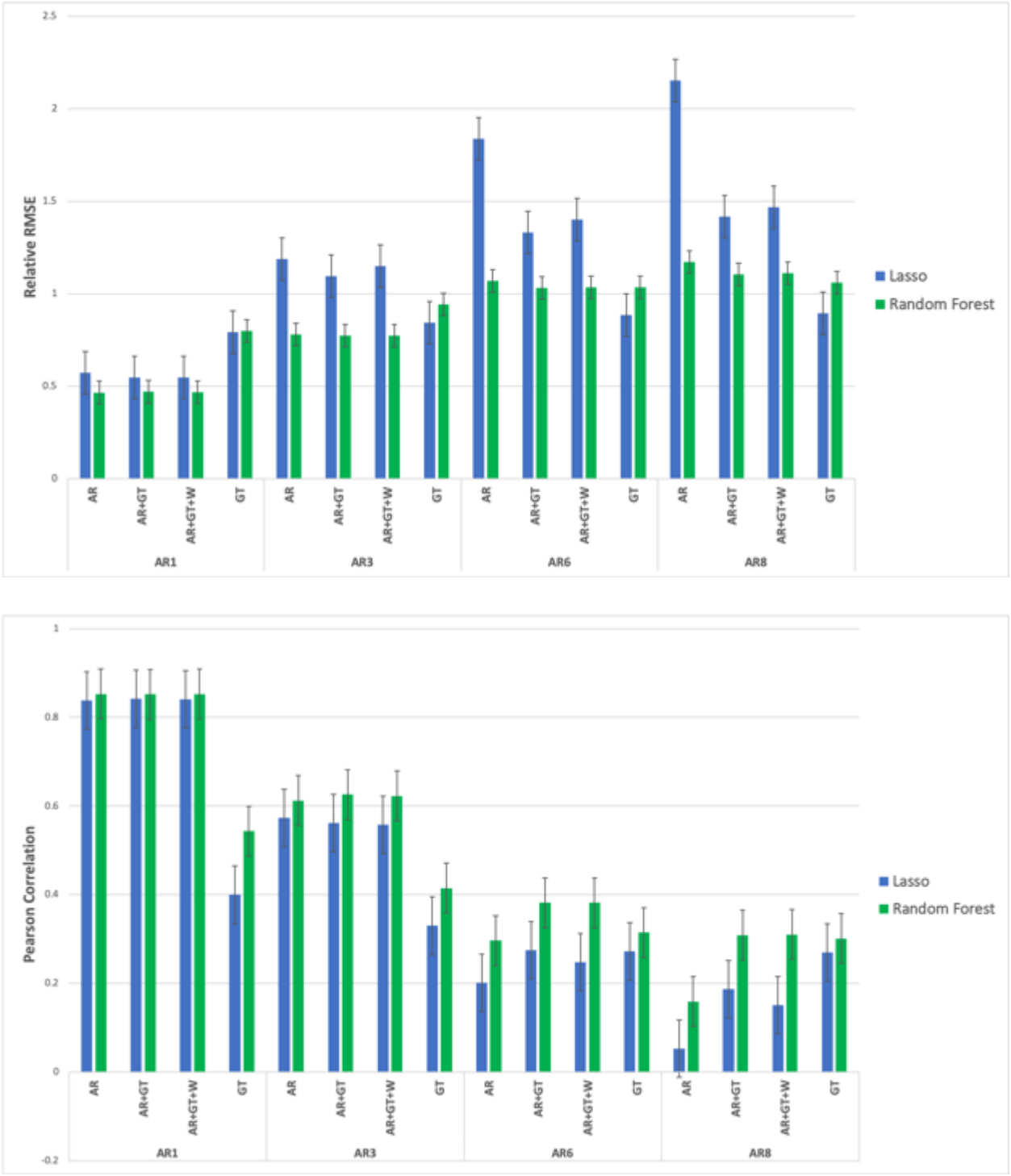


Figure 4: Change in relative importance of different predictors over time. Barra Mansa, Random Forest model with full feature set (autoregressive epidemiological data, google trends data, and weather data). Left: Assumed delay of 8 weeks in the availability of epidemiological data. Right: Assumed delay of 1 week in the availability of epidemiological data.

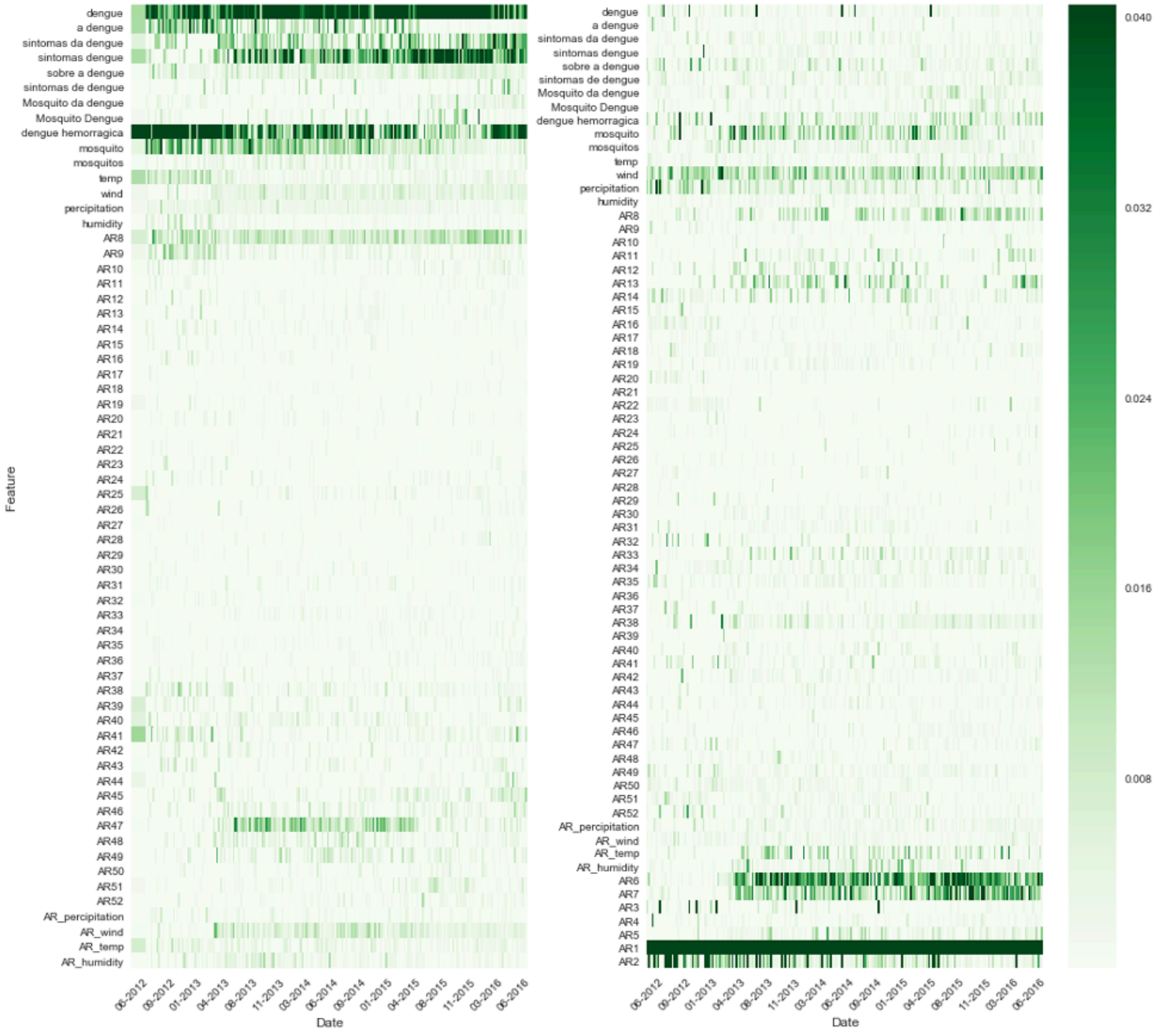
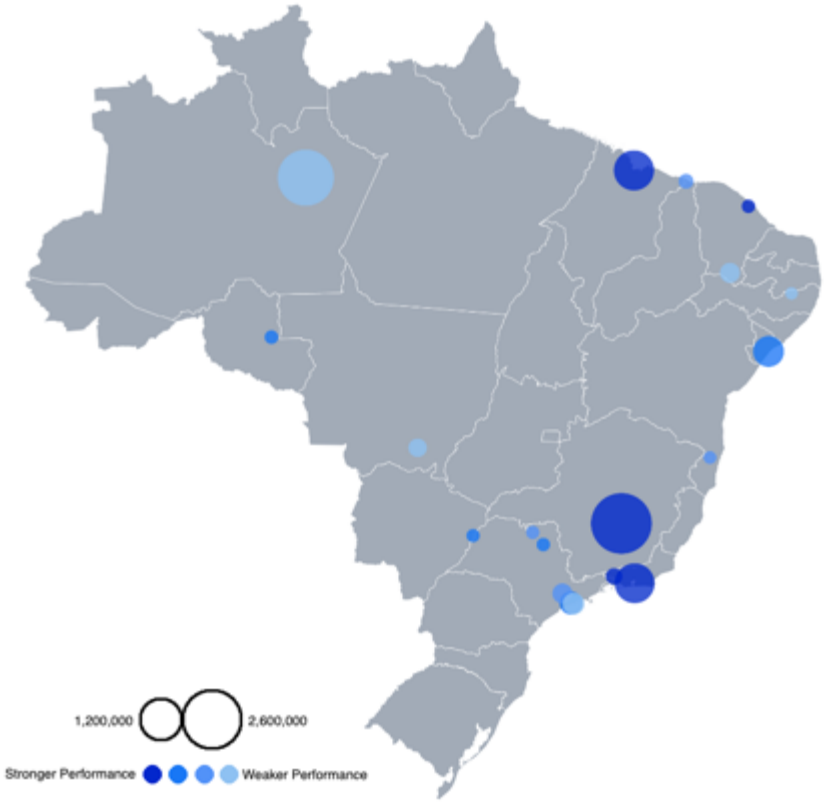
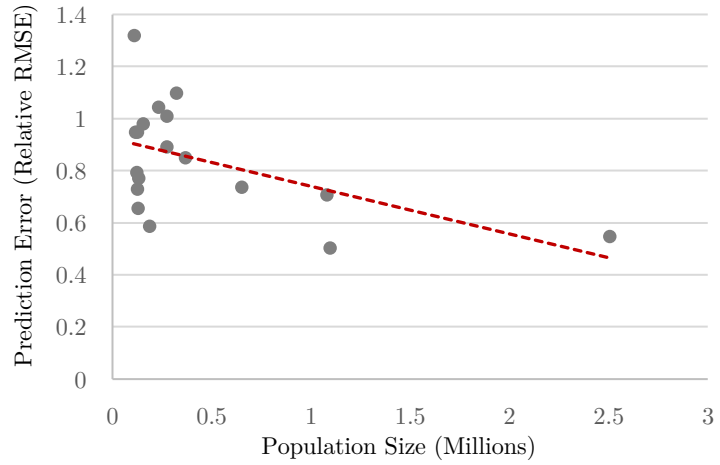
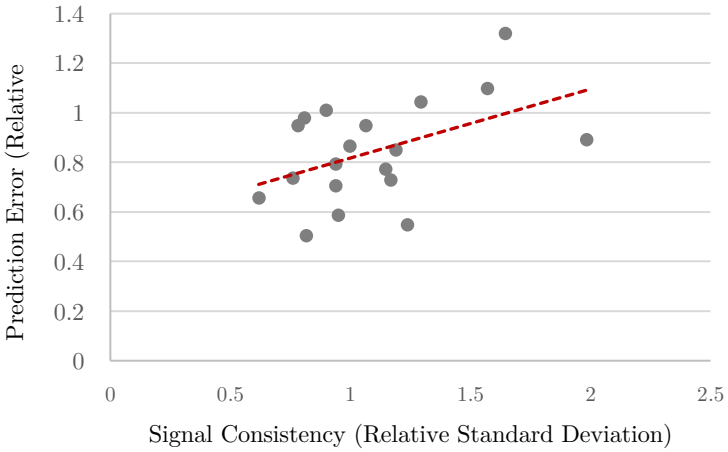
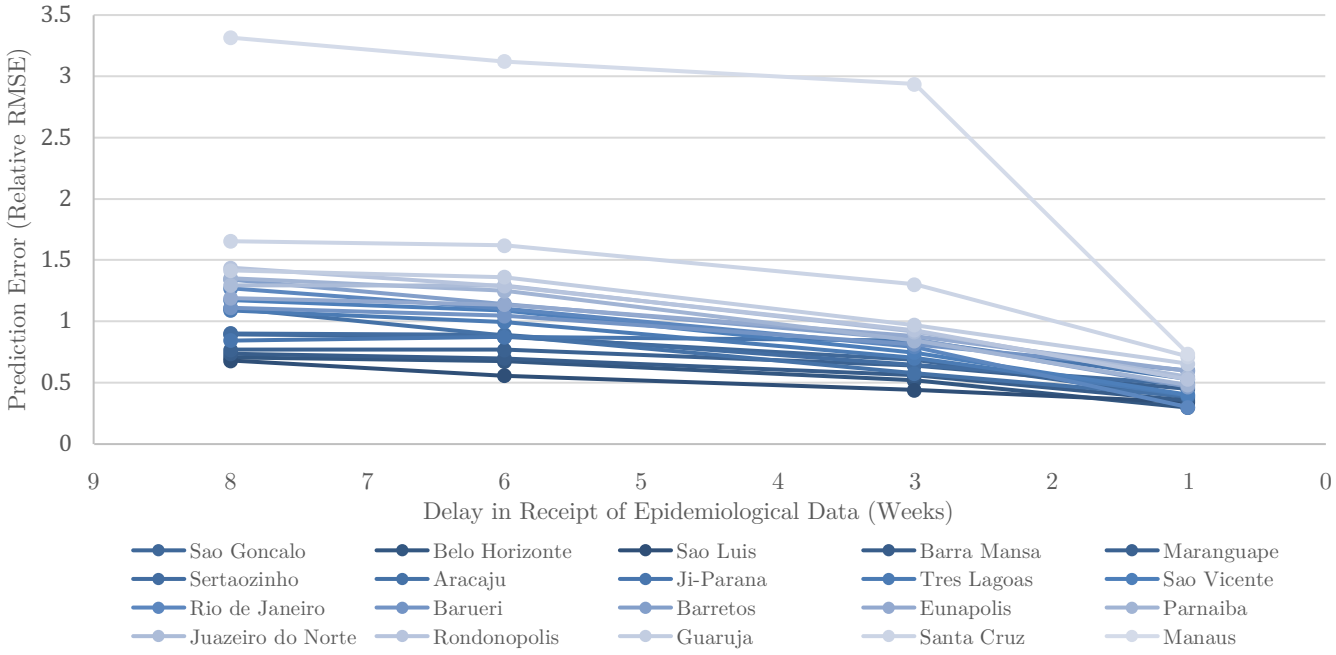


Figure 5: Determinants of success of nowcasting at the city level (random forest model, AR+GT feature set). **Top left.** City success, plotted on spatial map. **Top Right.** Effect of population size and dengue signal consistency on accuracy of predictions (averaged) **Bottom.** Prediction accuracy at different offsets









For the binary task, we generated retrospective out-of-sample predictions using both the Linear and Random Forest methodologies, between October 5 of 2012 and July 31 of 2017, for the 20 municipalities in Brazil. The total number of time intervals generated were 60 (3 per city). To measure our model's ability to predict an epidemic year, we utilized the standard definition of accuracy. We also measured the time difference Δ_t (in number of weeks) between t_p, the week when our models nowcasted a dengue epidemic, and t_e, the week in which the cumulative cases cross the epidemic threshold value. Δ_t is only measured for true positives (that is, in cases where t_p occurred earlier than t_e). These metrics are summarized in **Figure 6**. **Figure B in the SI Text** shows the distribution of epidemic and non-epidemic time intervals as a function of the epidemic threshold value. As the value of the epidemic threshold rises, the number of intervals classified as epidemic reduces, given the number of cumulative cases does not cross the threshold anymore.

Our results for the binary task show that our models are capable of successfully predicting epidemics, reaching accuracy values between .75 and .90, depending on the methodology and the type of information incorporated in the model. Lasso models achieve this with assumed delays in availability of "observed" epidemiological information of 5 to 7 weeks, whereas Random Forest-based models perform well with an assumed delay of up to 9 weeks. The choice of epidemic threshold does not affect these results.

Discussion

Despite the difficulties inherent to predictions at finer spatial resolutions, our results show that our models and methodological framework for nowcasting dengue, both LASSO-regression and random forest methods succeed at the city level and achieve accurate, actionable estimates. By accurately assessing suspected disease trends ahead of traditional disease surveillance systems, these nowcasts can enable decision-makers to better plan for and implement dengue mitigation policies. These include scheduling education and mosquito control programs, informing supply chain efforts for medical supplies, and warning of outbreaks that are expected to be particularly severe.

The conditions in which each methodology and data sources lead to more accurate performance vary. While the LASSO-based model has a slight edge at predictions that assume a longer delay in the availability of epidemiological information, the Random Forest-based model is more robust than the parametric model, and it tends to leverage the various data sources more effectively, with relatively little cost to interpretability. One possible reason for this is that tree-based models like random forests can capture non-linear relationships, which likely exist between at least a few of our features and dengue incidence counts.

The predictive power of the different sources of information (epidemiological data, Google search data, and weather) used in this study varied depending on the expected delays of epidemiological data reports. For predictions that assume very short delays in the availability of epidemiological data, short-term and seasonal autocorrelation were key to improving estimates and captured a substantial amount of dengue variability. For predictions that assume longer delays, the real-time Google search trends data captured the most substantial amount of dengue variability. This is intuitively to be expected: the longer the span of time that has elapsed since observed data was available, the more useful the real-time proxy of Google Search Trends data becomes. Google Search Trends data also proved to be extremely effective in cases of sudden outbreaks, particularly when the scale was large enough. Such was the case with Barueri, a city in the state of São Paulo, in which there was a sharp spike in the number of dengue cases in 2015, well above peak incidence in previous years. In

this instance, the feature set containing Google Search Trends data alone (GT) led to the most accurate performance at all time horizons, even when the assumed delay of epidemiological data was just a single week (see S1 text, appendix).

We find that long-term estimates tend to be more accurate when a city's population is larger and when past dengue incidence has been relatively regular (See Fig. 5, top right). As Google Search Trends data can only be collected at the state level in Brazil, it is reasonable that its relevance to nowcasts made at the city level is higher in cases where the examined city's population makes up a significant proportion of the state's population, as in Rio de Janeiro, for example (or in cases where different municipalities in the state exhibit similar dengue incidence patterns).

Finally, though in some cities with certain characteristics the models perform better than in others, they tend to adapt quite well to the specific patterns of each city (lags, peak size of outbreak, etc.) after a period of training on a city's past incidence data. Our framework contributes to the sparse but growing literature of infectious disease prediction models. Our results indicate that the lessons learned from dengue nowcasting in data-rich environments and at the country level can be generalized and tailored to track dengue in environments with significantly smaller populations, poorer data and a weaker disease signal. These insights can be leveraged towards future improvements in city-level nowcasting of infectious disease incidence.

Further Work

One epidemiological feature to be included as input in future models is dengue incidence in proximate cities. Recent work has shown that certain geographical regions of Brazil have become increasingly vulnerable to dengue as transport infrastructure and other means of transportation to them has improved. [35] Modelling this effect – for example, with cellular data, estimated volume of transportation, or simply with distance metrics – could improve estimates further, particularly for regions in which past observed case counts are less accurate or entirely unavailable. With the regression-based LASSO model, one naïve assumption that the relationship between the features and outcome variables is linear. This assumption is unlikely to be accurate (certainly across *all* variables), thus hampering model performance. But it could be that adding interaction and polynomial terms (which could then be narrowed down with a method like PCA) would improve LASSO performance, making it as robust as the Random Forest-based model, which does not assume linearity. Finally, one promising future direction is to design a composite model. This would take into account the finding that different feature sets, as well as the different underlying methodologies (LASSO and RF), led to the best performances in different cities and from different time lags. A composite model would incorporate these different sub-models and feature sets, and make use of them at the most fitting instances based on findings from the training data (for example, Google Search Trends data could be used as the feature set when making estimates that assume longer delays in the availability of observed case data).

Limitations

The weather data were produced at a naive resolution of 0.5 x 0.625 degrees, which works approximately to a ~50 square km grid cell. Attributing these data to a specific municipality, then, involved overlaying the rectangular grid of climate data onto a spatial file outlining municipal boundaries, and taking the weighted average of grid cells covering the municipality boundary. Thus, there are some data fluctuations that come from grid cells that partially cover the ocean, or different altitudes/mountains. More generally, the approximations in data modelled and assimilated from MERRA tend to lead to less noise than weather station data (precisely because it is modelled) – so there are tighter but potentially less accurate oscillations in the time series. Google

Search Trends data are only currently available at the state level in Brazil. Were they to be made available at finer spatial resolutions, such as the city level (as they currently are in the United States) it is expected that performance would improve. This effect is likely to be particularly significant when making predictions that assume greater delays in the availability of epidemiological data, in which the Google Search Trends data were the most important features driving the forecast.

Acknowledgements

MS was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM130668. GK, CB, and MS thank the Harvard Data Science Initiative for their support in the early stages of this project. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

Supporting Information

S1 Text. Supporting information text. This file includes: (1) Query terms used for Google Trends as Table A; (2) Characteristics for each of the 20 chosen cities in Table B; (3) Heatmaps of ARGO coefficients for each country/state as Figure A, B, C, D, and E. (PDF) (4). ; (5). ; (6). .

S2 Text. Measures of nowcasting performance across all models, features sets, and cities.

Author Contributions

Conceptualization: GK MS CB.

Data curation: GK SM TC.

Formal analysis: GK MS LC.

Funding acquisition: MS.

Investigation: GK MS.

Methodology: GK FL MS.

Project administration: MS.

Software:

Supervision: MS.

Validation: GK FL MS.

Visualization: GK.

Writing – original draft: GK FL MS.

All authors contributed and approved of the final version of the manuscript.

References

[1] World Health Organization. Dengue and severe dengue; 2016. http://www.who.int/mediacentre/factsheets/fs117/en/.

[2] Jane P. Messina, Simon I. Hay et al. "The current and future global distribution and population at risk of dengue." *Nature microbiology* (2019).

[3] World Health Organization and Special Programme for Research and Training in Tropical Diseases, World Health Organization E Department of Control of Neglected Tropical Diseases, World Health Organization. Dengue: guidelines for diagnosis, treatment, prevention and control. World Health Organization; 2009.

[4] Runge-Ranzinger S, Horstick O, Marx M, Kroeger A. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? Tropical Medicine & International Health. 2008; 13(8):1022–1041. https://doi.org/10.1111/j.1365-3156.2008.02112.x

[5] Madoff LC, Fisman DN, Kass-Hout T. A new approach to monitoring dengue activity. PLoS neglected tropical diseases. 2011; 5(5). https://doi.org/10.1371/journal.pntd.0001215 PMID: 21647309

[6] Chan M, Johansson MA. The incubation periods of dengue viruses. PloS one. 2012; 7(11):e50972. https://doi.org/10.1371/journal.pone.0050972 PMID: 23226436

[7] Centers for Disease Control and Prevention. Dengue; 2016. http://www.cdc.gov/dengue/.

[8] Ibarra AMS, Ryan SJ, Beltra´ n E, Mej´ıa R, Silva M, Muñoz A´. Dengue vector dynamics (Aedes aegypti) influenced by climate and social factors in Ecuador: implications for targeted control. PloS one. 2013; 8 (11):e78263. https://doi.org/10.1371/journal.pone.0078263

[9] Hii YL, Zhu H, Ng N, Ng LC, Rocklo¨v J. Forecast of dengue incidence using temperature and rainfall. PLoS Negl Trop Dis. 2012; 6(11):e1908. https://doi.org/10.1371/journal.pntd.0001908 PMID: 23209852

[10] Wongkoon S, Jaroensutasinee M, Jaroensutasinee K, et al. Distribution, seasonal variation & dengue transmission prediction in Sisaket, Thailand. Indian Journal of Medical Research. 2013; 138(3):347. PMID: 24135179

[11] Thai KTD, Anders KL. The role of climate variability and change in the transmission dynamics and geo- graphic distribution of dengue. Experimental Biology and Medicine. 2011; 236(8):944–954. https://doi. org/10.1258/ebm.2011.010402 PMID: 21737578

[12] Yang HM, Macoris MLG, Galvani KC, Andrighetti MTM, Wanderley DMV. Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. Epidemiology and infection. 2009; 137(08):1188–1202. https://doi.org/10.1017/S0950268809002052 PMID: 19192322

[13] Padmanabha H, Durham D, Correa F, Diuk-Wasser M, Galvani A. The interactive roles of Aedes aegypti super-production and human density in dengue transmission. PLoS Negl Trop Dis. 2012; 6(8): e1799. https://doi.org/10.1371/journal.pntd.0001799 PMID: 22953017

[14] Thammapalo S, Chongsuvivatwong V, Geater A, Dueravee M. Environmental factors and incidence of dengue fever and dengue haemorrhagic fever in an urban area, Southern Thailand. Epidemiology and Infection. 2008; 136(01):135–143. https://doi.org/10.1017/S0950268807008126 PMID: 17359563

[15] Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. Scientific Reports. 2016; 6. https://doi.org/10.1038/srep33707

[16] Promprou S, Jaroensutasinee M, Jaroensutasinee K. Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models. Dengue Bulletin. 2006; 30:99.
Luz PM, Mendes BVM, Codec¸o CT, Struchiner CJ, Galvani AP. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. The American journal of tropical medicine and hygiene. 2008; 79 (6):933–939. PMID: 19052308

[17] Choudhury ZM, Banu S, Islam AM. Forecasting dengue incidence in Dhaka, Bangladesh: A time series analysis. 2008;.

[18] Eastin MD, Delmelle E, Casas I, Wexler J, Self C. 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. 2014; 91(3):598–610. https://doi.org/10.4269/ajtmh. 13-0303 PMID: 24957546

[19] Johansson, Michael A., Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz et al. "An open challenge to advance probabilistic forecasting for dengue epidemics." *Proceedings of the National Academy of Sciences* 116, no. 48 (2019): 24268-24274.

[20] Codeco, C., F. Coelho, O. Cruz, S. Oliveira, T. Castro, and L. Bastos. "Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil." Revue d'Épidémiologie et de Santé Publique 66 (2018): S386.

[21] Bastos, Leonardo S., Theodoros Economou, Marcelo FC Gomes, Daniel AM Villela, Flavio C. Coelho, Oswaldo G. Cruz, Oliver Stoner, Trevor Bailey, and Claudia T. Codeço. "A modelling approach for correcting reporting delays in disease surveillance data." *Statistics in medicine* 38, no. 22 (2019): 4363-4377.

[22] Salmon, Maëlle, Dirk Schumacher, Klaus Stark, and Michael Höhle. "Bayesian outbreak detection in the presence of reporting delays." *Biometrical Journal* 57, no. 6 (2015): 1051-1067.

[23] van de Kassteele, Jan, Paul HC Eilers, and Jacco Wallinga. "Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing." *Epidemiology (Cambridge, Mass.)* 30, no. 5 (2019): 737.

[24] Bastos, Leonardo, Theodoros Economou, Marcelo Gomes, Daniel Villela, Trevor Bailey, and Claudia Codeço. "Modelling reporting delays for disease surveillance data." *arXiv preprint arXiv:1709.09150* (2017).

[25] de Almeida Marques-Toledo, Cecilia, Carolin Marlen Degener, Livia Vinhal, Giovanini Coelho, Wagner Meira, Claudia Torres Codeço, and Mauro Martins Teixeira. "Dengue prediction by the web: tweets are a useful tool for estimating and forecasting dengue at country and city level." *PLoS neglected tropical diseases* 11, no. 7 (2017): e0005729.

[26] Gelaro R, McCarty W, Suárez MJ, et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). J Clim 2017; 30: 5419–54.

[27] Ministry of Health, Epidemiological Report—dengue Fever (January to June, 2008). Available from (website in Portuguese): http://bvsms.saude.gov.br/bvs/publicacoes/informe_epidemiologico_dengue_janeiro_junho_2008.pdf

[28] https://publichealth.jmir.org/2019/2/e12214/

[29] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences. 2015; 112(47):14473–14478. https://doi. org/10.1073/pnas.1515373112

[30] Ho TK. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition 1995 Aug 14 (Vol. 1, pp. 278-282). IEEE.

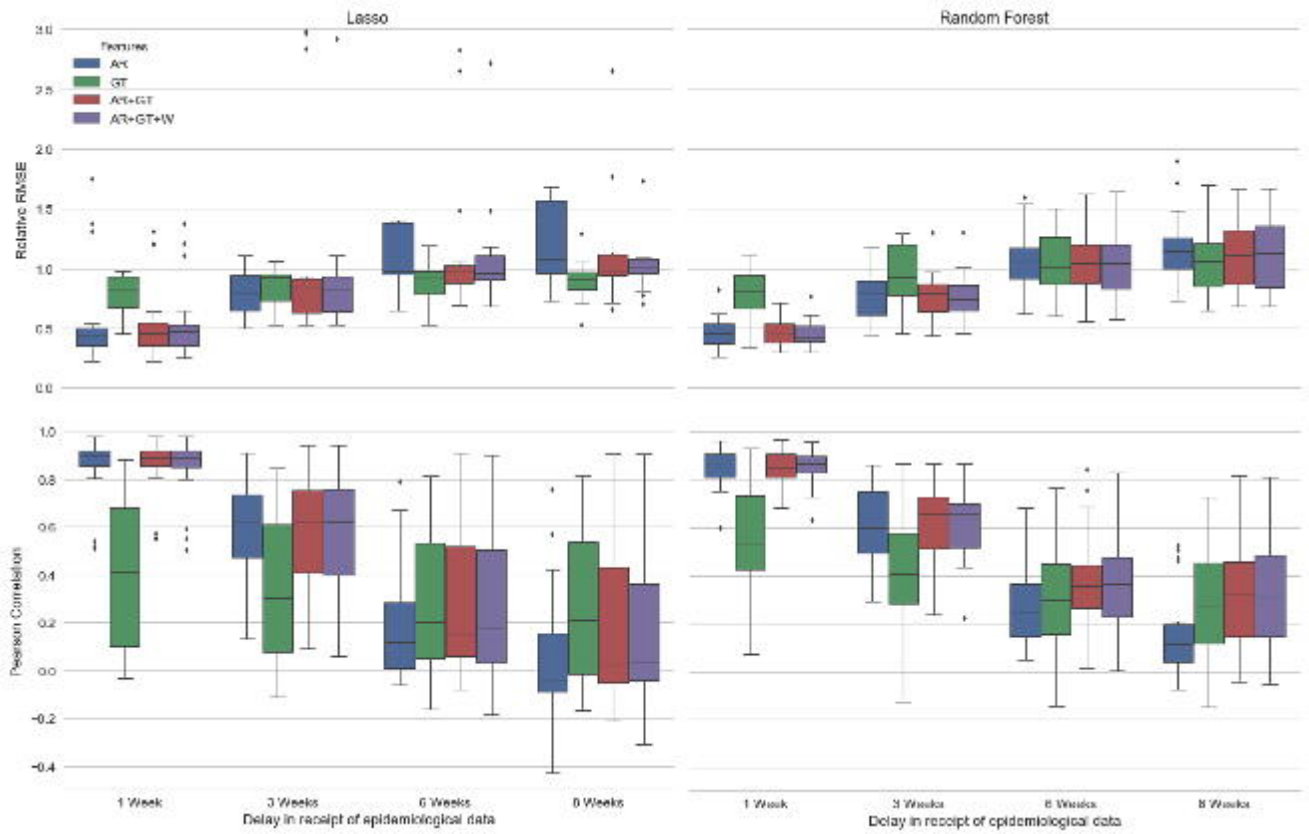
[31] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

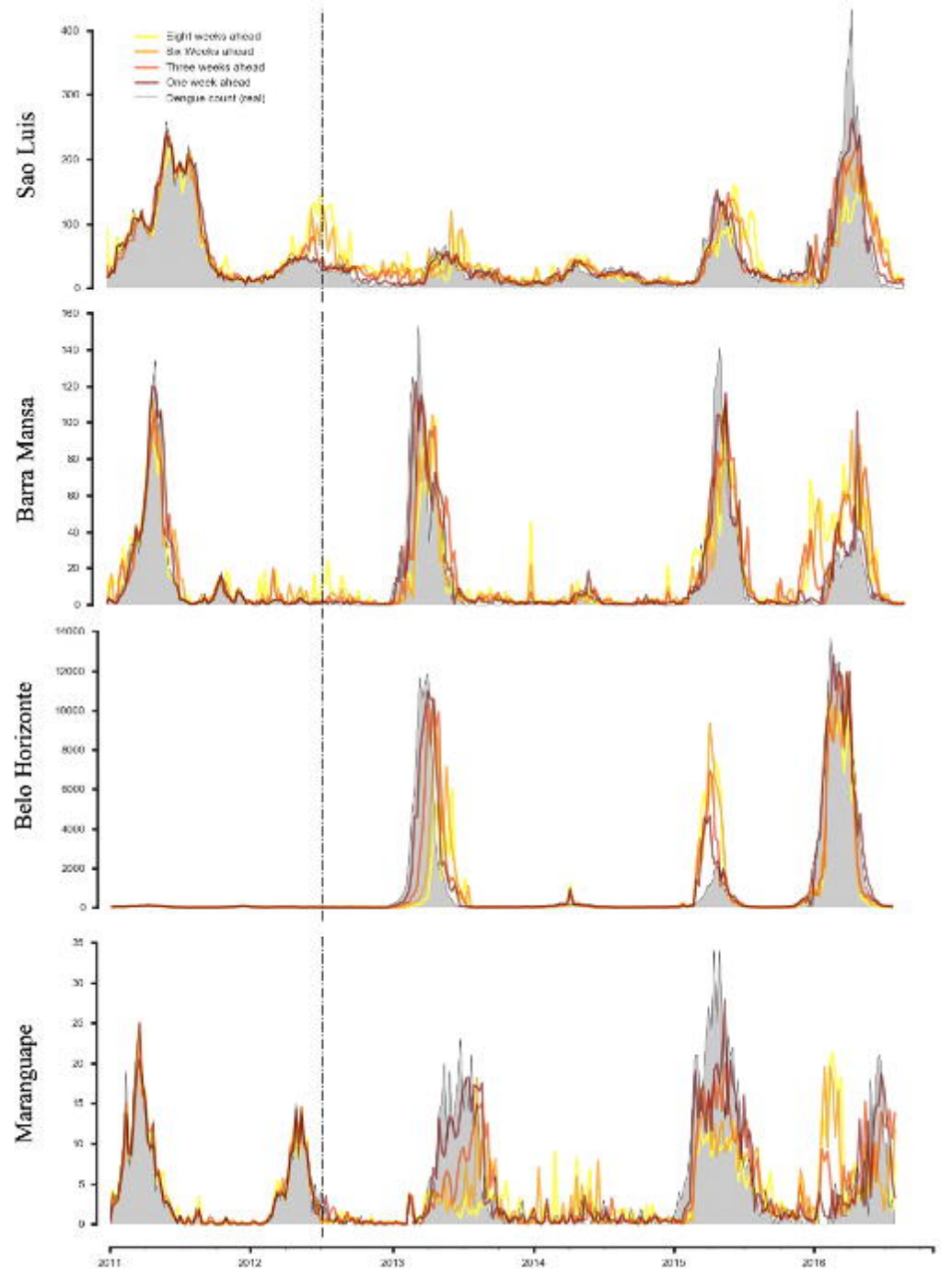
[32] Zhu R, Zeng D, Kosorok MR. Reinforcement learning trees. Journal of the American Statistical Association. 2015 Oct 2;110(512):1770-84.

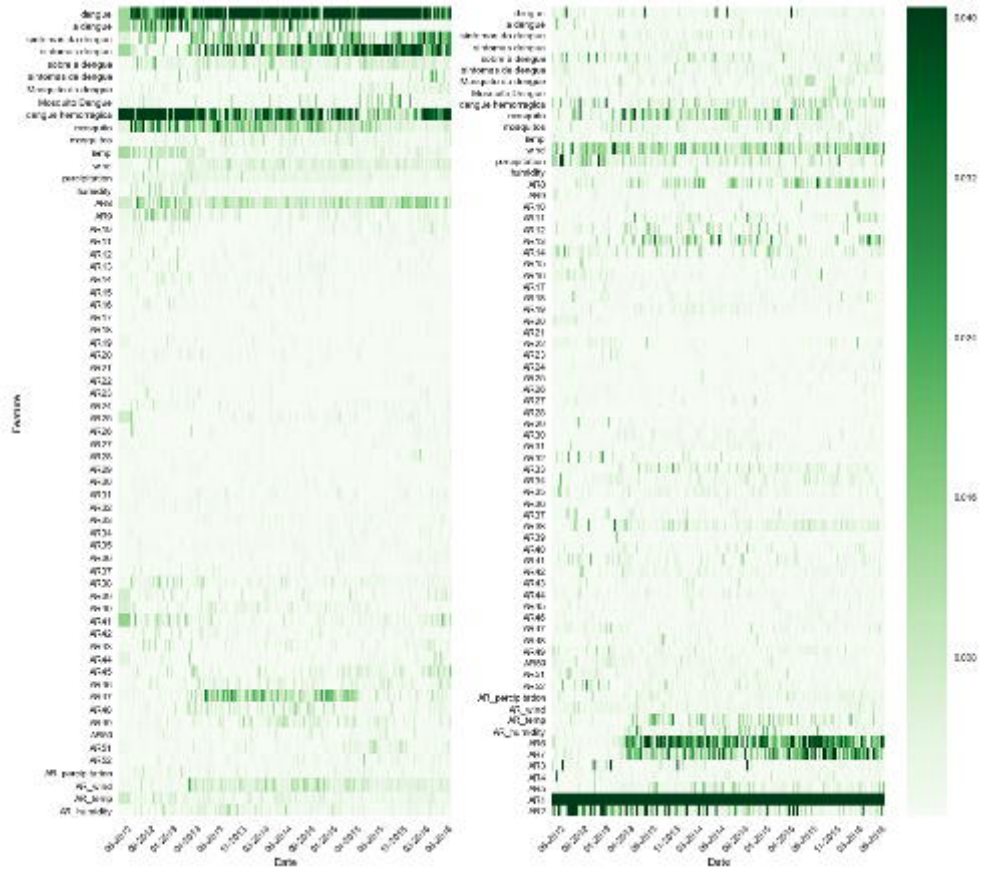
[33] Yang, S., Kou, S.C., Lu, F., Brownstein, J.S., Brooke, N. and Santillana, M., 2017. Advances in using Internet searches to track dengue. PLoS computational biology, 13(7), p.e1005607.

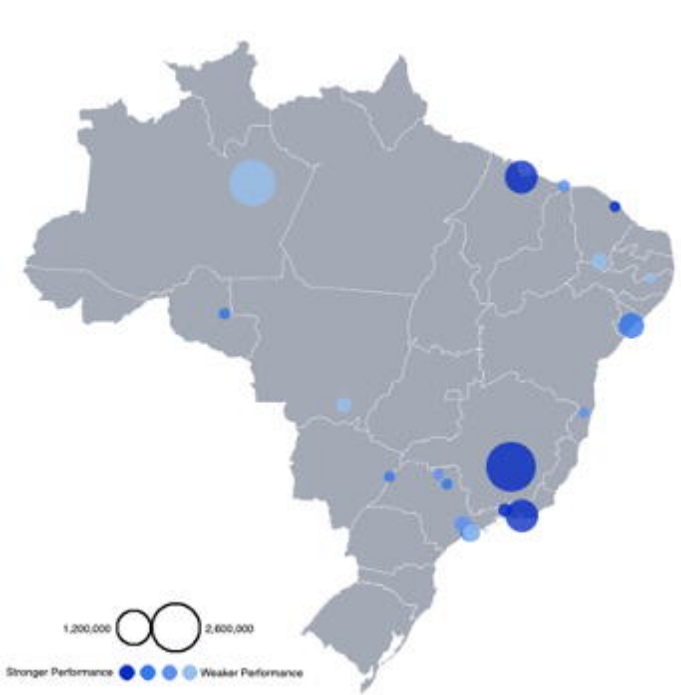
[34] Cortes, Fanny, Celina Maria Turchi Martelli, Ricardo Arraes de Alencar Ximenes, Ulisses Ramos Montarroyos, João Bosco Siqueira Junior, Oswaldo Gonçalves Cruz, Neal Alexander, and Wayner Vieira de Souza. "Time series analysis of dengue surveillance data in two Brazilian cities." *Acta tropica* 182 (2018): 190-197.

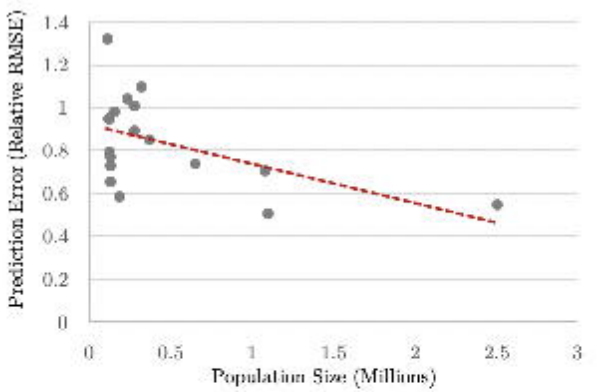
[35] Lana, Raquel Martins, Marcelo Ferreira da Costa Gomes, Tiago Franca Melo de Lima, Nildimar Alves Honorio, and Cláudia Torres Codeço. "The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: a network-based analysis." *PLoS neglected tropical diseases* 11, no. 11 (2017): e0006070.

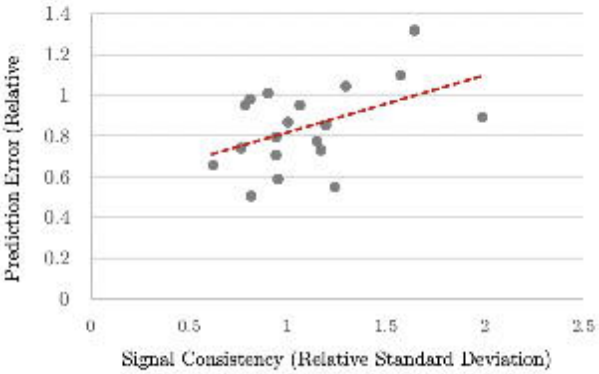

Lasso
Random Forest
Features
AR
GT
AR+GT
AR+GT+W
Relative RMSE
Pearson Correlation
1 Week
3 Weeks
6 Weeks
8 Weeks
Delay in receipt of epidemiological data

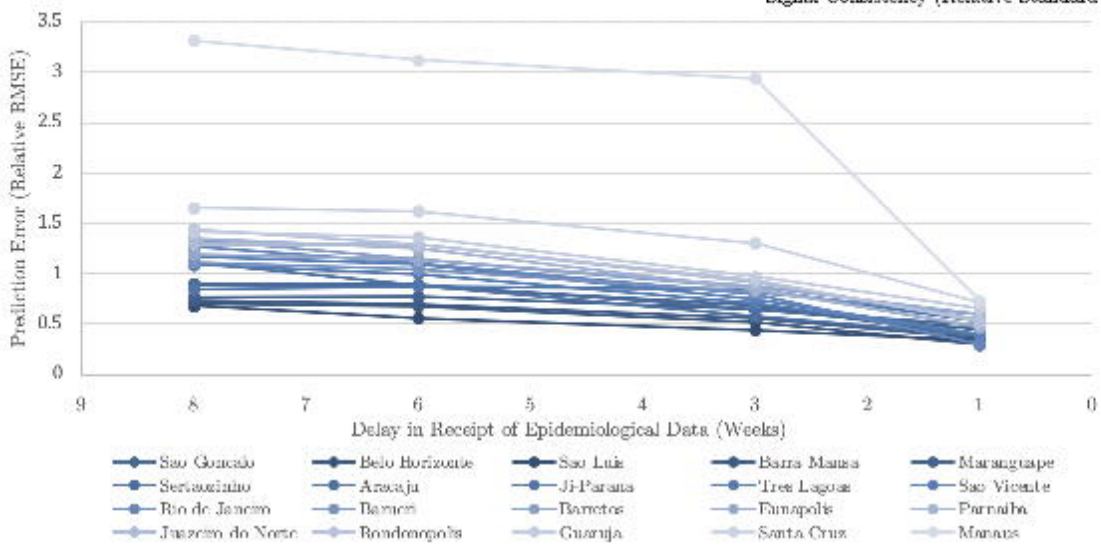

Eight weeks ahead
Six Weeks ahead
Three weeks ahead
One week ahead
Dengue count (real)
Sao Luis
Barra Mansa
Belo Horizonte
Maranguape
2011
2012
2013
2014
2015
2016

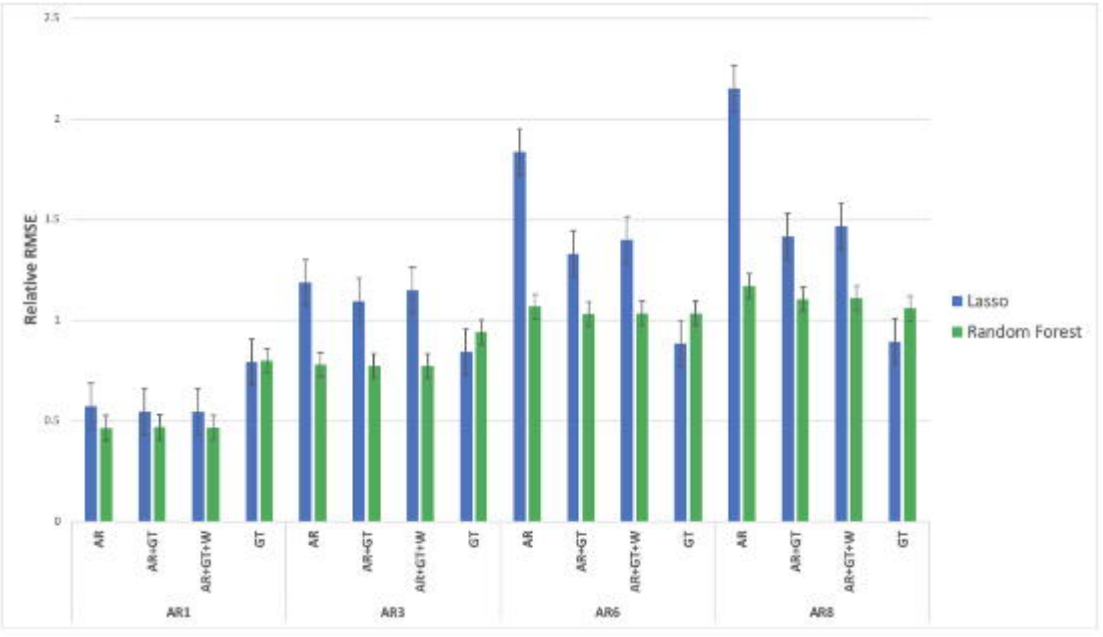

Famílias
Data
0.040
0.032
0.024
0.016
0.008

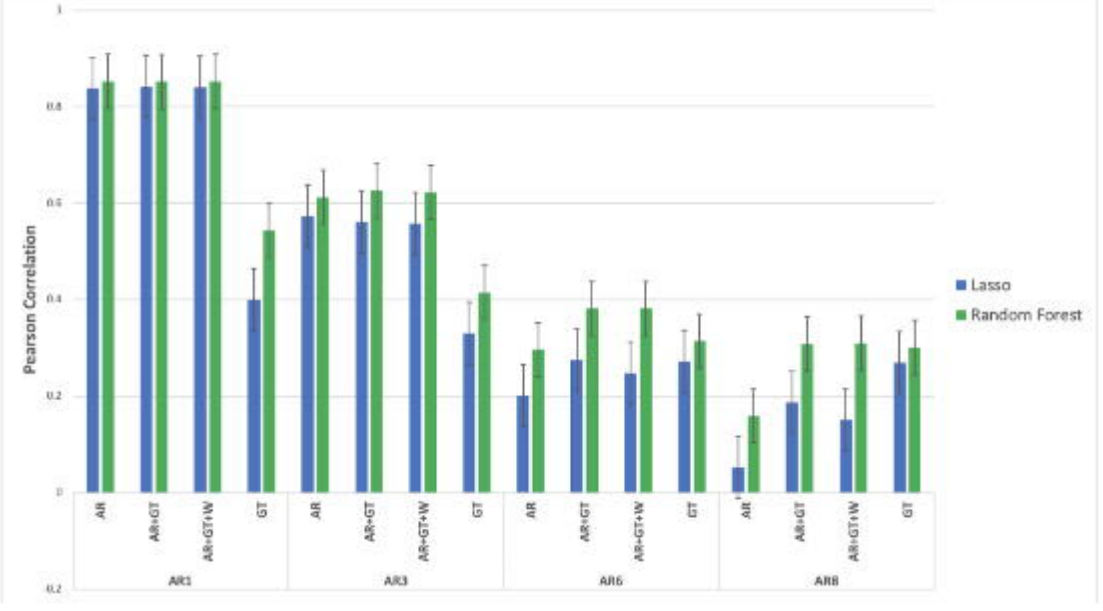

1,200,000
2,600,000
Stronger Performance
Weaker Performance


Prediction Error (Relative RMSE)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
Population Size (Millions)


Prediction Error (Relative
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
Signal Consistency (Relative Standard Deviation)


Prediction Error (Relative RMSE)
3.5
3
2.5
2
1.5
1
0.5
0
9
8
7
6
5
4
3
2
1
0
Delay in Receipt of Epidemiological Data (Weeks)
Sao Goncalo
Belo Horizonte
Sao Luis
Barra Mansa
Maranguape
Serrastiano
Aracaju
Ji-Parana
Tres Lagoas
Sao Vicente
Rio de Janeiro
Barueri
Barretos
Fumapolis
Parnaiba
Juazeiro do Norte
Rondonopolis
Guaruja
Santa Cruz
Manaus


Relative RMSE
2.5
2
1.5
1
0.5
0
AR
AR+GT
AR+GT+W
GT
AR1
AR3
AR6
AR8
Lasso
Random Forest


Pearson Correlation
1
0.8
0.6
0.4
0.2
0
AR
AR+GT
AR+GT+W
GT
AR1
AR3
AR6
AR8
Lasso
Random Forest