bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burdens at the national sub-national scale in Colombia

Naizhuo Zhao[1,2], Katia Charland[3], Mabel Carabali[4], Elaine Nsoesie[5], Mathieu Maher-Giroux[4], Erin Rees[6], Mengru Yuan[4], Cesar Garcia Balaguera[7], Gloria Jaramillo Ramirez[7], Kate Zinszer[3,8*]

1. Institute of Land Resource Management, School of Humanities and Law, Northeastern University, Shenyang, Liaoning, China.

2. Division of Clinical Epidemiology, McGill University Health Centre, Montréal, QC, Canada.

3. Centre de recherche en santé publique, Université de Montréal et CIUSSS du Centre-Sud-de-l'Île-de-Montréal, Montréal, Quebéc, Canada

4. Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Quebec, Canada

5. Department of Global Health, Boston University, Boston, Massachusetts, USA

6. Public Health Risk Sciences Division, National Microbiology Laboratory, Public Health Agency of Canada, Saint-Hyacinthe, Québec, Canada

7. Faculty of Medicine, Cooperative University of Colombia, Villavicencio, Meta, Colombia

8. Department of Preventive and Social Medicine, University of Montreal, Montréal, Québec, Canada.

*Corresponding author: Kate Zinszer
E-mail: kate.zinszer@umontreal.ca

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract:

The robust estimate and forecast capability of random forests (RF) has been widely recognized, however this ensemble machine learning method has not been widely used in mosquito-borne disease forecasting. In this study, two sets of RF models were developed for the national and departmental levels in Colombia to predict weekly dengue cases at 12-weeks ahead. A national model based on artificial neural networks (ANN) was also developed and used as a comparator to the RF models. The various predictors included historic dengue cases, satellite-derived estimates for vegetation, precipitation, and air temperature, population counts, income inequality, and education. Our RF model trained on the national data was more accurate for department-specific weekly dengue cases estimation compared to a local model trained only on the department's data. Additionally, the forecast errors of the national RF model were smaller to those of the national ANN model and were increased with the forecast horizon increasing from one-week ahead (mean absolute error, MAE: 5.80; root mean squared error, RMSE: 11.10) to 12-weeks ahead (MAE: 13.38; RMSE: 26.82). There was considerable variation in the relative importance of predictors dependent on forecast horizon. The environmental and meteorological predictors were relatively important for short-term dengue forecast horizons while socio-demographic predictors were relevant for longer-term forecast horizons. This study showed the potential of RF in dengue forecasting with also demonstrating the feasibility of using a national model to forecast at finer spatial scales. Furthermore, sociodemographic predictors are important to include to capture longer-term trends in dengue.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Author summary:

Dengue virus has the highest disease burden of all mosquito-borne viral diseases, infecting 390 million people annually in 128 countries. Forecasting is an important warning mechanism that can help with proactive planning and response for clinical and public health services. In this study, we compare two different machine learning approaches to dengue forecasting: random forest (RF) and neural networks (NN). National and local (departmental-level) models were compared and used to predict dengue cases in the future. The results showed that the counts of future dengue cases were more accurately estimated by RF than by NN. It was also shown that environmental and meteorological predictors were more important for forecast accuracy for shorter-term forecasts while socio-demographic predictors were more important for longer-term forecasts. Finally, the national model applied to local data was more accurate in dengue forecasting compared to the local model. This research contributes to the field of disease forecasting and highlights different considerations for future forecasting studies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Dengue virus is most prevalent of the mosquito-borne viral diseases, infecting 390 million people annually in 128 countries with four different virus serotypes [1]. Rising incidence and large-scale outbreaks are largely due to inadequate living conditions, naïve populations, global trade and population mobility, climate change, and the adaptive nature of the principal mosquito vectors *Aedes aegypti* and *Aedes albopictus* [2, 3]. The direct and indirect costs of dengue are substantial and impose enormous burdens on low- and middle-income tropical countries, with a global estimate of US$8.9 billion in costs per year [4].

Human and financial costs of dengue can be alleviated when response systems, such as intervention strategies, health care services, supply chain management, receive timely warnings of future cases through forecasting models [5]. A number of dengue forecasting models have been developed and these models can be generally classified into two methodological categories: time-series and machine learning [6, 7]. The majority of existing dengue forecasting models used time-series methods and typically Autoregressive Integrated Moving Average (ARIMA), in which lagged meteorological factors (e.g. temperature and precipitation) act as covariates in conjunction with historical dengue data for one- to 12-week ahead forecasting [8-13]. Many studies reported that conventional time-series models such as ARIMA are insufficient to meet complex forecasting requirements [14-16], as multiple trends and outliers present in the time-series reduce the forecasting accuracy [17].

In the last two decades, machine learning (ML) methods have been used in many disciplines, such as geography, environment, and epidemiology, to yield meaningful findings from highly heterogeneous data. Machine learning statistical regression methods are promising approaches for disease forecasting as they facilitate the inclusion of a large number of correlated

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

105 variables, enable the modeling of complex interactions between variables, and can fit complex
106 models without strong parametric assumptions that are often untestable in traditional statistical
107 approaches [18, 19]. Decision trees, support vector machine, artificial neural network, K-nearest
108 neighbor, gradient boosting, and naive Bayes are frequently used ML approaches in dengue-
109 forecasting studies [7, 20-23]. Compared to the above ML methods, random forests (RF) have
110 shown to be more accurate in forecasting given its ability to overcome the common problem of
111 over-fitting through the use of bootstrap aggregation [24-28].

112 Random forests have been used to forecast dengue risk in several countries including
113 Costa Rica [29], Philippines [30, 31] Pakistan [32], Peru and Puerto Rico [33]. However, time or
114 seasonal variables were not always included in the models nor were sociodemographic
115 predictors, which have been found to improve forecast accuracy in HIV [34] and Ebola [35]
116 epidemic models. Furthermore, dengue models, regardless of the use of the time series or ML
117 approaches, have been developed for predicting dengue cases in individual administrative areas
118 such in a city or a province [9-12, 20-23]. Universal dengue prediction models that are effective
119 across different administrative regions remain absent.

120 Historically, Colombia is one of the countries most affected by dengue, with the *Aedes*
121 mosquito being widely distributed throughout all departments at elevations below 2,000 meters
122 [36, 37]. The objective of this study was to evaluate the potential of RF forecasting models at the
123 department and national level in Colombia. We compared the accuracy of the department and
124 national RF models to understand the feasibility of using a national model to predict dengue
125 cases for individual departments. We also compared errors of the national RF models with those
126 of Artificial Neural Network (ANN), another classic and widely used ML approach. Finally, we
127 estimated the change in importance of different predictors according to forecast horizon.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data and methods

Data

Various data were used to develop the forecasting models, which included: dengue cases from surveillance data, environmental indicators from remoting sensing data, and sociodemographic indicators such as population, income inequity, and education coverage (Table 1). The dengue case surveillance data were extracted from an electronic platform, SIVIGILA, created by the Colombia national surveillance program and was available at the department level. The national surveillance program receives weekly reports from all public health facilities that provide services to cases of dengue. The dengue cases reported by SIVIGILA were a mixture of probable and laboratory confirmation. Laboratory confirmation for dengue is based on a positive result from antigen, antibody, or virus detection and/or isolation [38]. Confirmation of probable cases is largely based on clinical diagnosis plus at least one serological positive immunoglobulin M test or an epidemiological link to a confirmed case 14 days prior to symptom onset.

Table 1. Summary of study indicators and data sources

Indicator	Source	Temporal granularity	Format
Dengue cases	The national surveillance program in Colombia	Weekly	Tabular
Rainfall	CMORPH precipitation data from NOAA's CPC	Daily	Gridded
EVI	MOD13C1 from NASA's LP DAAC	16-day	Gridded
Temperature	MOD11C2 from NASA's LP DAAC	8-day	Gridded
Population	Colombian National Administrative Department of Statistics	Yearly	Tabular
Gini index	Colombian National Administrative Department of Statistics	Yearly	Tabular
Education coverage	Colombian National Administrative Department of Statistics	Yearly	Tabular

CMORPH, Climate Prediction Center morphing method; CPC, Climate Prediction Center; EVI, enhanced vegetation index; LP DAAC, Land Processes Distributed Active Archive Center; NASA, National Aeronautics and Space Administration; NOAA, National Oceanic and Atmospheric Administration.

Precipitation, air temperature, and land cover type have been shown to be three important determinants of *Aedes* mosquito abundance and are often used as predictors in dengue

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

forecasting [9, 11, 21, 39]. In this study, precipitation data was obtained from the CMORPH (Climate Prediction Center morphing method) daily estimated precipitation dataset [40]. The land surface temperatures were extracted from the MODIS Terra Land Surface Temperature 8-day image products. Enhanced vegetation index (EVI) estimates were obtained from the MODIS Terra Vegetation Indices 16-Day image products. Several studies have shown that socio-demographic factors may influence dengue transmission and incidence as significantly as environmental factors [41-43]. Given this, we included population, Gini index (a measure of income inequity), and education coverage as potential predictors, which were retrieved from the Colombian National Administrative Department of Statistics. The study was approved by the Sciences and Health Ethical Committee of the University of Montreal (CERSES-19-018D), and all data were provided at the aggregate level and are publicly available.

Random forests

Random forests (RF) is an ensemble decision tree approach [44]. A decision tree is a simple representation for classification in which each internal node corresponds to a test on an attribute, each branch represents an outcome of a test, and each leaf (i.e. terminal node) holds a class label. Decision trees can also be used for regression when the target or outcome variable is continuous. Bootstrap aggregation, commonly known as bagging, is the most distinctive technique used in RF and bagging requires training each decision tree with a randomly selected subsample of the entire training datasets.

Data preprocessing

To ensure a consistent temporal granularity with the outcome variable, the daily precipitation data were aggregated to a weekly frequency. The 8-day land surface temperature and the 16-day EVI data were resampled to a weekly frequency using a spline interpolation [45].

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We assigned a given department the same population, Gini index, and education coverage values for all weeks within the same calendar year.

The archipelago of San Andrés, Providencia, and Santa Catalina (commonly known as *San Andrés y Providencia*) is a department consisting of two island groups and 775 km away from mainland Colombia. Due to the frequent cloud contamination over the small island areas, it was not possible to have high-quality MODIS images products for weekly temperature or EVI value estimation. Vaupés department had only 30 confirmed dengue cases scattered in 24 weeks during 2014 to 2018. Thus, the departments of San Andrés y Providencia and Vaupés were excluded from this study, and data from the other 30 departments were used to train our models.

Weekly dengue data from 2014-2017 was used to train the RF models and the data from 2018 was used to evaluate the models. To simulate 'real life' forecasting, we did not include the 2018 data for the socio-demographic variables given that they are only produced annually whereas the remote sensing data are more readily available. Exponential smoothing approach based on historical (2010-2017) time-series data to estimate the values for 2018.

Development of RF models

We first developed RF models for each department (referred to as local level). Let the "current" week be k and the number of confirmed dengue cases be y. Referring to the RF streamflow forecasting model developed by Papacharalampous and Tyralis [46], we used the numbers of current and previous 11 weeks dengue cases (i.e. y_k, y_{k-1}, ..., y_{k-10}, y_{k-11}) of a department to predict one-week ahead dengue cases (i.e. y_{k+1}) for each department. The current and previous 11 weeks of rainfall, land surface temperature, EVI, population, Gini index, and education coverage were also included as predictors. These values were selected as previous studies demonstrated that the optimal lags of meteorological variables used for dengue

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

196 forecasting are usually not larger than 12 weeks [47-52]. In addition, the ordinal number of the
197 forecast week (1–52 for the year of 2015, 2016, 2017, and 2018 and 1–53 for 2014) as well as
198 year (2014–2018) were treated as two predictor variables to account for seasonality and long-
199 term changing trend of dengue occurrence [53,54].

200 We then developed RF models at the national scale. To train a national-scale RF model
201 for forecasting *n*-week ahead dengue cases (where $n \leq 12$), we used the same predictor and target
202 variables as those used in the local *n*-week ahead forecasting models. The difference between the
203 local and the national models was that the local *n*-week ahead models were trained using 209-*n*
204 (209 =53+52+52+52) samples while the national model was trained using 6270-30*n* [i.e. (209-n)
205 ×30] samples.

206 ***Model evaluation***

207 Model accuracy was evaluated and compared by two metrics: mean absolute error (MAE)
208 and root mean squared error (RMSE). The MAEs and RMSEs reported in this study were
209 calculated by the actual and the predicted numbers of dengue cases for the 52 weeks in 2018.
210 The accuracy comparison was performed at the local (department) and national scales. When the
211 comparison for an *n*-week ahead prediction was conducted at the national scale, the predicted
212 numbers of dengue cases by the 30 local RF models were additively combined and compared
213 with the actual national values to calculate one MAE and one RMSE. When the comparison was
214 implemented at the local scale, the national RF model was applied to each one of the 30
215 departments and then the predicted values were compared with the actual numbers of dengue
216 cases to compute 30 individual MAEs and 30 individual RMSEs.

217 Artificial Neural Network (ANN) is an early ML approach and has been previously used
218 to predict dengue cases [7, 20, 21, 23]. We developed ANN models at the national scale and

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

compared their prediction accuracy with that of the RF models. The ANN was composed of one input layer, one hidden layer, and one output layer. The ANN models had the same 53 predictor variables as the RF models, resulting in 53 neurons in the input layer and one neuron in the output layer. The number of neurons in the hidden layer was determined by iterative attempts until the prediction accuracy cannot be further improved [55]. In this study, the optimal number of neurons in hidden layer varied by forecasting horizon and ranged between 38 to 50.

Percentage of increased mean squared error (%IncMSE) is a robust and informative indicator to quantitatively evaluate the importance of predictor variables in a random forests model [56]. Percentage of increased mean squared error indicates the increase in the mean squared error (MSE) of prediction as a result of an independent variable being randomly shuffled while maintaining the other independent variables as unchanged [44]. A larger %IncMSE of a predictor variable suggests greater importance of the variable on the model's overall forecast accuracy and the %IncMSE was calculated for each predictor in each RF model.

Results

An exceptionally large dengue outbreak occurred in Colombia during the study period. The counts of confirmed dengue cases reached more than 2,500 per week by the end of 2015 and the outbreak ended mid-year in 2016. Following this outbreak, the yearly dengue case peaks were drastically reduced in 2016 and 2017 but began increasing again in 2018 (Fig1).

Fig 1. The weekly total counts of confirmed dengue cases over Colombia for 2014-2018 (A) and the predicted counts of dengue cases by the national one-, two-, four-, eight-, and twelve-week ahead models for 2018 (B). See Fig S1 for the predicted counts of dengue cases for all week ahead models.

For any of the n-week ahead ($n \leq 12$) forecasts, the performance of the national model was better than that of the local model, demonstrated by the smaller overall MAE and RMSE (Table

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2). Moreover, in most cases, a department's dengue cases were more accurately predicted by the national model than the local model (Fig 2). The errors of the national random forests model were mainly derived from under-estimation of cases which coincided with dramatic increases in cases towards the end of 2018. As expected, the under-estimation was more pronounced when predictions were made over a longer time period.

Table 2. Comparison of accuracy between the local and the national models

n-week ahead	Local RF model		National RF model		National ANN model	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	8.01	14.74	5.80	11.10	8.81	13.02
2	9.10	17.05	6.76	13.43	10.59	15.86
3	10.14	19.24	7.64	14.79	11.72	18.17
4	11.05	21.67	8.52	16.20	12.70	19.89
5	12.14	24.14	9.23	17.76	14.13	22.15
6	12.86	25.80	10.08	18.89	15.12	27.15
7	13.50	27.25	10.77	20.55	16.73	28.07
8	13.94	28.04	11.46	22.19	18.06	28.27
9	14.36	29.09	11.95	23.50	18.62	28.99
10	14.67	29.66	12.51	24.82	20.34	32.38
11	14.91	30.15	12.93	25.90	21.25	33.14
12	15.21	30.66	13.38	26.82	21.93	33.89

MAE, mean absolute error; RMSE, root mean squared error; RF, random forests; ANN, artificial neural network.

Fig 2. Accuracy comparison between the local and the national random forests models at the department scale for the one-week ahead, four-week ahead, eight-week ahead, and twelve-week ahead predictions with RMSE for 2018. See Fig S2-S4 in the supporting information on MAE and for all week ahead models.

The overall RMSE of the ANN model developed at the national scale was smaller than that of the local RF model at forecasting horizons of 5 weeks or less (Table 2). The RMSE grew for the ANN model with longer forecasting horizons compared to the local RF model. The MAE of the ANN model was consistently larger than that of the local RF model for each forecasting

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

279 horizon. The RMSE and MAE of the national RF model were smaller than those of the national
280 ANN model at any forecasting horizon.

281 The relative importance of different predictor variables in the national RF model was
282 varied (Table 3). Firstly, "current" and "near current" past dengue data were extremely important
283 in predicting occurrence of dengue in the near future (e.g. one- to three-weeks ahead). However,
284 with the predicted week increasingly further away from the "current" week, the importance of
285 historical dengue data decreased while the "current" week of dengue cases remained one of the
286 top three most important predictors in predicting the future dengue cases. Secondly, the
287 environmental (EVI) and the meteorological predictors (rainfall and temperature) were more
288 important than the socio-demographic predictors when dengue cases were predicted in the near
289 future (one- to three-weeks ahead). Yet, with the predicted week increasingly far away from the
290 "current" week, the three socio-demographic covariates (education, population, and Gini index)
291 became increasingly important. Finally, the week predictor, which accounted for the seasonal
292 pattern of dengue, was important across all forecasting horizons but relatively smaller in
293 importance with smaller forecasting horizons (i.e. $n \leq 4$)

294
295
296

Table 3. The top ten most important predictor variables for predicting dengue cases in the national models, ordered from the largest to the smallest %IncMSEs

Rank	1	2	3	4	5	6	7	8	9	10
1-week ahead	$Dengue_{k}$	$Dengue_{k-1}$	$Dengue_{k-2}$	$Dengue_{k-3}$	Week	$Dengue_{k-4}$	EVI_{k-11}	$Temperature_{k-11}$	EVI_{k-10}	EVI_{k-8}
2-week ahead	$Dengue_{k}$	$Dengue_{k-1}$	Week	$Dengue_{k-2}$	$Dengue_{k-3}$	$Temperature_{k-11}$	$Dengue_{k-4}$	EVI_{k-7}	EVI_{k-5}	EVI_{k-8}
3-week ahead	$Dengue_{k}$	$Dengue_{k-1}$	Week	$Dengue_{k-2}$	EVI_{k-8}	EVI_{k-10}	$Temperature_{k-10}$	Education	$Dengue_{k-3}$	$Dengue_{k-4}$
4-week ahead	$Dengue_{k}$	Week	$Dengue_{k-1}$	Education	$Dengue_{k-2}$	$Temperature_{k-9}$	EVI_{k-8}	$Temperature_{k-11}$	EVI_{k-7}	$Dengue_{k-3}$
5-week ahead	$Dengue_{k}$	Week	$Dengue_{k-1}$	Education	$Dengue_{k-2}$	EVI_{k-10}	$Temperature_{k-8}$	$Temperature_{k}$	Gini	EVI_{k-9}
6-week ahead	$Dengue_{k}$	Week	$Dengue_{k-1}$	Education	Population	Year	$Dengue_{k-2}$	EVI_{k-8}	EVI_{k-9}	EVI_{k-10}
7-week ahead	$Dengue_{k}$	Week	Education	$Dengue_{k-1}$	Year	$Dengue_{k-2}$	Population	Gini	EVI_{k-10}	EVI_{k-9}
8-week ahead	$Dengue_{k}$	Week	Population	Education	$Dengue_{k-1}$	Year	$Temperature_{k-11}$	$Temperature_{k-5}$	$Dengue_{k-2}$	Gini
9-week ahead	$Dengue_{k}$	Week	Population	Education	Year	$Dengue_{k-1}$	$Temperature_{k-11}$	$Dengue_{k-11}$	Gini	$Temperature_{k-3}$
10-week ahead	$Dengue_{k}$	Week	Year	Education	Population	$Dengue_{k-1}$	Gini	$Dengue_{k-11}$	$Temperature_{k-4}$	$Dengue_{k-2}$
11-week ahead	Year	Week	$Dengue_{k}$	Population	Education	Gini	$Dengue_{k-1}$	$Temperature_{k-11}$	$Dengue_{k-10}$	$Temperature_{k-4}$
12-week ahead	Population	Year	$Dengue_{k}$	Week	Education	Gini	$Dengue_{k-11}$	$Dengue_{k-1}$	$Dengue_{k-10}$	$Temperature_{k-10}$

Dengue indicates historical dengue cases and EVI denotes enhanced vegetation index. %IncMSE, percentage of increased mean squared error.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In the current study, we developed a national model to predict counts of dengue cases across different departments of Colombia and found that for the majority of departments, the national model more accurately forecasted future dengue cases at the department level compared to the local model. This result indicates the similarity in importance of dengue drivers across different administrative regions of Colombia. Random forests is an unsupervised tree-based regression approach requiring a relatively large training sample for the repeated splitting of the dataset into separate branches, and thus the national model trained by a larger dataset had higher prediction accuracy compared to the local models. The national and the local models performed poorly in departments of Guainía and Vichada. The small population and consequently, the low counts of dengue cases resulted in the relatively large errors in the two departments.

We found that the meteorological and environmental variables were more important for prediction accuracy at smaller forecasting horizons compared to the socio-demographic variables, with socio-demographics being more important at larger forecasting horizons. This is likely due to the influence of meteorological and environmental conditions on *Aedes* mosquitoes and the lag effects are usually between 1 to 4 weeks for temperature and precipitation [57-59]. Poor quality housing and sanitation management with high population density are key risk factors for dengue transmission [60, 61], and are closely related to education and poverty [62, 63]. These results demonstrate the complimentary nature of these different groups of predictor variables and the importance of their inclusion in dengue forecasting models.

We used ANN models as comparators to our RF models. Artificial Neural Networks are brain-inspired systems that are intended to imitate the way that human learn. Theoretically, more complex correlations between predictor and target variables can be discerned by deeper (i.e.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

more hidden layers) networks [64]. However, ANN cannot handle the problem of vanishing gradient which results in the failure of improving accuracy of ANN models by adding more hidden layers. Additionally, it is easy for ANN to suffer from over-fitting which leads to a network developed by a training dataset failing to predict the other observations accurately. In this study, the number of neurons in the hidden layer was required to be changed with each forecast horizon, demonstrating the poor universality of the ANN models. By contrast, RF solves the problem of over-fitting with the use of bootstrap aggregation. Hyperparameters (e.g. the number of decision trees) in RF are easy to be set and the RF models showed better universality for different forecast horizons.

Despite the strengths of our study, an important limitation with our RF approach is that the considerable dependence on the current week of dengue leads the model to generate lags for forecasting rapid changes in dengue. Including a predictor of mosquito abundance from an entomological surveillance program may reduce such time lag errors [65]. However, this type of data is often difficult to obtain at the national level with sufficient temporal and spatial granularity. Additionally, RF, as a non-parametric black-box approach, cannot intuitively display quantitative relationships between the count of dengue cases and the heterogeneous predictor variables, although it is able to more flexibly and accurately model the possibly complex non-linear and non-additive relationships among the variables. A more severe limitation of the RF model is the fact that RF cannot obtain values beyond the range of the variable in the training dataset. If an unprecedented dengue outbreak occurred in future, under-estimations will occur inevitably using the RF approach.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conclusions

This study highlights the potential of RF for dengue forecasting and also demonstrates the benefits of including socio-demographic predictors. Our findings also found that a national model, on average, performed better compared to the local models. Future studies should consider the inclusion of other arboviruses as predictors, such as chikungunya and Zika as well as examine the importance of other socio-economic factors. In addition, other promising ML methods should be tested including recurrent neural networks, which inherently account for time, have the ability to deal with a vanishing gradient, and are able to capture complicated non-linear and non-additive relationships between predictor and target variables [66].

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Lambrechts L, Scott TW, Gubler DJ. Consequences of the expanding global distribution of Aedes albopictus for dengue virus transmission. PLoS Neglected Tropical Diseases 2010; 4(5): e646.

[2] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL et.al. The global distribution and burden of dengue. Nature 2013; 496:504-507.

[3] Morin CW, Comrie AC, Ernst K. Climate and dengue transmission: evidence and implications. Environmental Health Perspectives 2013; 121(11-12): 1264.

[4] Shepard DS, Undurraga EA, Hallasa YA, Stanaway JD. The global economic burden of dengue: a systematic analysis. Lancet Infectious Diseases 2016; 16:935-941.

[5] Soyiri IN, Reidpath DD. An overview of health forecasting. Environmental Health and Preventive Medicine 2013; 18(1):1–9.

[6] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. PLoS Neglected Tropical Diseases 2012; 6(5):e1648.

[7] Gambhir S, Malik SK, Kumar Y, The diagnosis of dengue disease: An evaluation of three machine learning approaches. International Journal of Healthcare Information Systems and Informatics 2018; 13:1-19.

[8] Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K, Tong S, Climate change and dengue: a critical and systematic review of quantitative modelling approaches. BMC Infectious Diseases 2014; 14:167.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[9] Gharbi M, Quenel P, Gustave J, Cassadou S, Ruche GL, Girdary L, et al. Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. BMC Infectious Diseases 2011; 11:166.

[10] Hu W, Clements A, Williams G, Tong S, Dengue fever and El Niño/Southern Oscillation in Queensland, Australia: a time series predictive model. Occupational & Environmental Medicine 2010; 67:307-311.

[11] Dom NC, Hassan AA, Latif ZA, Ismail R, Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malasia. Asian Pacific Journal of Tropical Disease 2013; 3:352-361.

[12] Cortes F, Turchi Martelli CM, Arraes de Alencar Ximenes R, Montarroyos UR, Siqueira Junior JB, Gonçalves Cruz O, et al. Time series analysis of dengue surveillance data in two Brazilian cities. Acta Tropica. 2018; 182:190–7.

[13] 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; 6:33707.

[14] Niu M, Wang Y, Sun S, Li Y, A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term $PM_{2.5}$ concentration forecasting. Atmospheric Environment 2016; 134:168-180.

[15] Chen M-Y, Chen B-T, A hybrid fuzzy time series model based on granular computing for stock price forecasting. Information Sciences 2015; 294:227-241.

[16] Wang P, Zhang H, Qin Z, Zhang G, A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting. Atmospheric Pollution Research 2017; 8: 850-860.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

412 [17] Zhao N, Liu Y, Vanos JK, Cao G, Day-of-week and seasonal patterns of $PM_{2.5}$
413 concentrations over the United States: Time-series analyses using the Prophet procedure.
414 Atmospheric Environment 2018; 192:116-127.
415 [18] Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the
416 author). Statistical Science 2001; 16(3): 199-231.
417 [19] Murphy KP. Machine Learning: a probabilistic perspective. MIT Press, 2012.
418 [20] Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast
419 model using machine learning: A case study in China. PLoS Neglected Tropical Diseases 2017;
420 11:e0005973.
421 [21] Scavuzzo JM, Trucco F, Espinosa M, Tauro CB, Abril M, Scavuzzo CM, et al. Modeling
422 dengue vector population using remotely sensed data and machine learning. Acta Tropica 2018;
423 185:167-175.
424 [22] Althouse BM, Ng YY, Cummings DAT, Prediction of dengue incidence using serach query
425 surveillance. PLoS Neglected Tropical Diseases 2011; 5:e1258.
426 [23] Laureano-Rosario, AE, Duncvan AP, Mendez-Lazaro, PA, Garcia-Rejon JE, Gomez-Carro
427 S, Farfan-Ale J, et al. Application of artificial neural networks for dengue fever outbreak
428 predictions in the northwest coast of Yucatan, Mexico and San Juan, Puerto Rico. Tropical
429 Medicine and Infectious Disease 2018; 3:5.
430 [24] Raczko E, Zagajewski B, Comparison of support vector machine, random forest and neural
431 network classifiers for tree species classification on airborne hyperspectral APEX images.
432 European Journal of Remote Sensing 2017; 50:144-154.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

433 [25] Meyer H, Kulhnlein M, Appelhans T, Nauss T, Comparison of four machine learning
434 algorithms for their applicability in satellite-based optical rainfall retrievals. Atmospheric
435 Research 2016; 169:424-433.
436 [26] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M, Machine
437 learning predictive models for mineral prospectivity: An evaluation of neural networks, random
438 forest, regression trees and support vector machines. Ore Geology Reviews 2015; 71:804-818.
439 [27] Statnikov A, Wang L, Aliferis CF, A comprehensive comparison of random forests and
440 support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008;
441 9:319.
442 [28] Nsoesie EO, Beckman R, Marathe M, Lewis B, Prediction of an epidemic curve: A
443 supervised classification approach. Statistical communications in infectious diseases.
444 2011; 3(1):5.
445 [29] Vasquez P, Loria A, Sanchez F, Barboza LA, Climate-driven statistical models as effective
446 predictors of local dengue incidence in Costa Rica: A generalized additive model and random
447 forest approach. arXiv 2019; 1907.13095.
448 [30] Olmoguez ILG, Catindig MAC, Amongos MFL, Lazan AF, Developing a dengue
449 forecasting model: A case study in Iligan city. International Journal of Advanced Computer
450 Science and Applications 2019; 10(9):281–286.
451 [31] Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K, Machine
452 learning methods reveal the temporal pattern of dengue incidence using meteorological factors in
453 metropolitan Manila, Philippines. BMC Infectious Diseases 2018; 18:183.
454 [32] Rehman NA, Kalyanaraman S, Ahmad T, Pervaiz F, Saif U, Subramanian L, Fine-grained
455 dengue forecasting using telephone triage services. Science Advances 2016; 2(7): e1501215.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

456 [33] Freeze J, Erraguntla M, Verma A, Data integration and predictive analysis system for
457 disease prophylaxis: Incorporating dengue fever forecasts. Proceedings of the 51st Hawaii
458 International Conference on System Science 2018; 913-922.
459 [34] Dinh L, Chowell G, Rothenberg R, Growth scaling for the early dynamics of HIV/AIDS
460 epidemics in Brazil and the influence of socio-demographic factors. Journal of Theoretical
461 Biology 2018; 442:79-86.
462 [35] Chretien J-P, Riley S, George DB, Mathematical modeling of the West Aftica Ebola
463 epidemic. eLIFE 2015; 4:e09186.
464 [36] Cardona-Ospina JA, Villamil-Gómez WE, Jimenez-Canizales CE, Castañeda-Hernández
465 DM, Rodríguez-Morales AJ. Estimating the burden of disease and the economic cost attributable
466 to chikungunya, Colombia, 2014. Transactions of the Royal Society of Tropical Medicine and
467 Hygiene 2015; 109(12):793–802.
468 [37] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological trends of dengue disease
469 in Colombia (2000-2011): a systematic review. PLoS Neglected Tropical Diseases 2015; 9(3):
470 e0003499.
471 [38] Ospina Martinez ML, Martinez Duran ME, Pacheco García OE, Bonilla HQ, Pérez NT.,
472 Protocolo de vigilancia en salud pública enfermedad por virus Zika. PRO-R02.056. Bogota
473 (Colombia): Instituto Nacional de Salud, 2017. Available from:
474 http://bvs.minsa.gob.pe/local/MINSA/3449.pdf (last accessed December 16, 2019).
475 [39] Beketov MA, Yurchenko YA, Belevich OE, Liess M, What environmental factors are
476 important determinants of structure, species richness, and abundance of mosquito assemblages?
477 Journal of Medical Entomology 2010; 47:129-139.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

478 [40] Joyce RJ, CMORPH: A method that produces global precipitation estimates from passive
479 microwave and infrared data at high spatial and temporal resolution. Journal of
480 Hydrometeorology 2004; 5:487-503.
481 [41] Koyadun S, Butraporn P, Kittayapong P, Ecologic and sociodemographic risk determinants
482 for dengue transmission in urban areas in Thailand. Interdisciplinary Perspectives on Infectious
483 Diseases 2012; 2012:907494.
484 [42] Reiter P, Climate change and mosquito-borne disease. Environmental Health Perspectives
485 2001; 109(supplement 1):141-161.
486 [43] Soghaier MA, Himatt S, Osman KE, Okoued SI, Seidahmed OE, Beatty ME, et al., Cross-
487 sectional community-based study of the socio-demographic factors associated with the
488 prevalence of dengue in the eastern part of Sudan in 2011. BMC Public Health 2015; 15:558.
489 [44] Breiman L, Random forests. Machine learning 2001; 45(1):5-32.
490 [45] Hulme M, New M. Dependence of large-scale precipitation climatologies on temporal and
491 spatial sampling. Journal of Climate, 1997; 10:1099–1113,
492 [46] Papacharalampous GA, Tyralis H, Evaluation of random forests and prophet for daily
493 streamflow forecasting. Advances in Geosciences 2018; 45:201-208.
494 [47] Lu L, Lin H, Tian L, Yang W, Sun J, Liu Q, Time series analysis of dengue fever and
495 weather in Guangzhou, China, BMC Public Health 2009; 9:395.
496 [48] Chen S-C. Liao C-M, Chio C-P, Chou H-H, You S-H, Cheng Y-H, lagged temperature
497 effect with mosquito transmission potential explains dengue variability in southern Taiwan:
498 Insights from a statistical analysis. Science of The Total Environment 2010; 408(19):469-4075.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

499 [49] Cheong YL, Burkart K, Leitao PJ, Lakes T, Assessing weather effects on dengue disease in
500 Malaysia, International Journal of Environmental Research and Public Health 2013;
501 10(12):6319-6334.
502 [50] Chang K, Chen, C-D, Shih C-M, Lee T-C, Wu M-T, Wu D-C, et al., Time-lagging interplay
503 effect and excess risk of meteorological/mosquito parameters and petrochemical gas explosion
504 on dengue incidence. Scientific reports 2016; 6:35028.
505 [51] Chen Y, Ong JHY, Rajarethinam J, Yap G, Ng LC, Cook AR. Neighbourhood level real-
506 time forecasting of dengue cases in tropical urban Singapore. BMC Medicine 2018;16(1):129.
507 [52] Eastin MD, Delmelle E, Casas I, Wexler J, Self C, Intra-and interseasonal autoregressive
508 prediction of dengue outbreaks using local weather and regional climate for a tropical
509 environment in Colombia. The American Journal of Tropical Medicine and Hygiene 2014;
510 91(3):598-610.
511 [53] Bostan N, Javed S, Amen N, Eqani SAMAS, Tahir F, Bokhari H, Dengue fever virus in
512 Pakistan: effects of seasonal pattern and temperature change on distribution of vector and virus.
513 Reviews in Medical Virology 2017; 27(1):e1899.
514 [54] Oidtman RJ, Lai S, Huang Z, Yang J, Siraj AS, Reiner RC, et al., Inter-annual variation in
515 seasonal dengue epidemics driven by multiple interacting factors in Guangzhou, China, Nature
516 Communications 2019; 10:1148.
517 [55] Peng Z, Letu H, Wang T, Shi C, Zhao C, Tana G, Zhao N, Dai T, Tang R, Shang H, Shi J,
518 Chen L. Estimation of shortwave solar radiation using the artificial neural network from
519 Himawari-8 satellite imagery over China. Journal of Quantitative Spectroscopy and Radiative
520 Transfer 2020; 240: 106672.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

521 [56] Liu Y, Cao G, Zhao N, Mulligan K, Ye X. Improve ground-level $PM_{2.5}$ concentration
522 mapping using a random forests-based geostatistical approach. Environmental Pollution 2018;
523 235: 272-282.
524 [57] Grziwotz F, Strauß JF, Hsieh C-h, Telschow A. Empirical dynamic modelling identifies
525 different responses of *Aedes Polynesiensis* subpopulations to natural environmental variables.
526 Scientific Reports 2018; 8: 16768.
527 [58] da Cruz Ferreira DA, Degener CM, de Almeida Marques-Toledo, C, Bendati MM, Fetzer
528 LO, Teixeira CP, Eiras AE. Meteorological variables and mosquito monitoring are good
529 predictors for infestation trends of Aedes aegypti, the vector of dengue, chikungunya and
530 Zika. Parasites Vectors 2017; 10: 78.
531 [59] Manica M, Filipponi F, D'Alessandro A, Screti A, Neteler M, Rosà R, et al. Spatial and
532 Temporal Hot Spots of *Aedes albopictus* Abundance inside and outside a South European
533 Metropolitan Area. PLoS Neglected Tropical Diseases 2016; 10(6): e0004758.
534 [60] Mulligan K, Dixon J, Sinn C-L J, Elliott SJ. Is dengue a disease of poverty? A systematic
535 review. Pathogens and Global Health 2015; 109(1): 10-18.
536 [61] Tapia-Conyer R, Méndez-Galván JF, Gallardo-Rincón H. The growing burden of dengue in
537 Latin America. Journal of Clinical Virology 2009; 46: S3-S6.
538 [62] Adams EA, Boateng GO, Amoyaw JA. Socioeconomic and demographic predictors of
539 potable water and sanitation access in Ghana. Social Indicators Research 2016; 126(2): 673-687.
540 [63] de Janvry A, Sadoulet E. Growth, poverty, and inequality in Latin America: A causal
541 analysis, 1970-94. The review of Income and Wealth 2000; 46(3): 267-287.
542 [64] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep
543 learning applications and challenges in big data analytics. Journal of Big Data 2015; 2:1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906297; this version posted January 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

544 [65] Ong J, Liu X, Rajarethinam J, Kok, SY, Liang S, Tang, CS, et al., Mapping dengue risk in

545 Singapore using random forest. PLoS Neglected Tropical Diseases 2018; 12(6):e0006587.

546 [66] Williams RJ, Zipser D, A learning algorithm for continually running fully recurrent neural

547 networks. Neural Computation 1989; 1(2):270-280.

Support Information Legends

Fig S1. The weekly total counts of confirmed dengue cases over Colombia for 2014-2018 (A) and the predicted counts of dengue cases by the national model for one to twelve-week ahead for 2018 (B).

Fig S2. Accuracy comparison between the local and the national random forests models at the department scale for the one-week ahead, four-week ahead, eight-week ahead, and twelve-week ahead predictions with MAE for 2018.

Fig S3. Accuracy comparison between the local and the national random forests models at the department scale for one to twelve-week ahead predictions with RSME for 2018.

Fig S4. Accuracy comparison between the local and the national random forests models at the department scale for one to twelve-week ahead predictions with MAE for 2018.

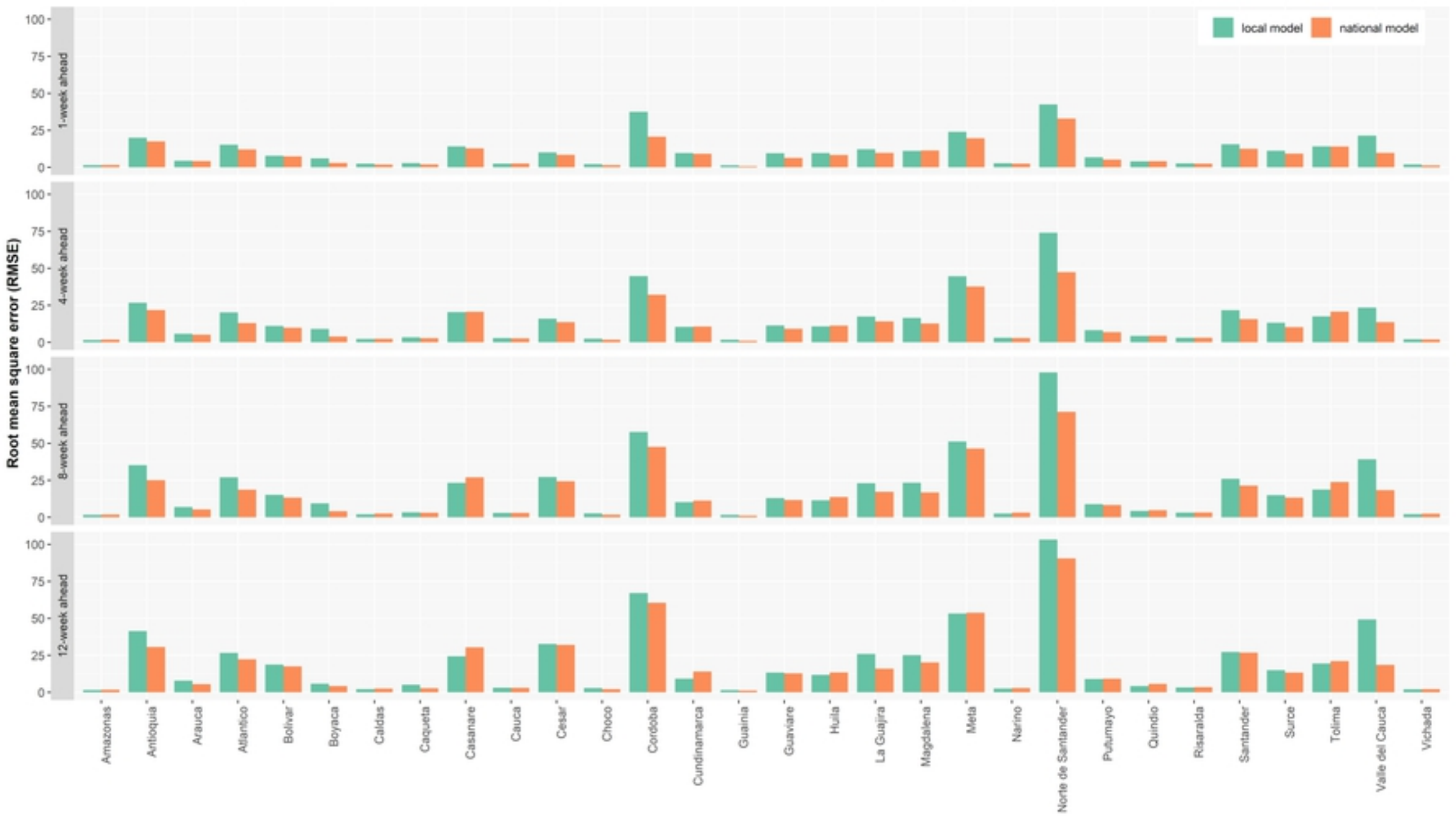
local model
national model
Root mean square error (RMSE)
1-week ahead
4-week ahead
8-week ahead
12-week ahead
0
25
50
75
100
Amazonas
Antioquia
Arauca
Atlantico
Bolivar
Boyaca
Caldas
Caqueta
Casanare
Cauca
Cesar
Choco
Cordoba
Cundinamarca
Guainia
Guaviare
Huila
La Guajira
Magdalena
Meta
Narino
Norte de Santander
Putumayo
Quindio
Risaralda
Santander
Sucre
Tolima
Valle del Cauca
Vichada

Figure 2

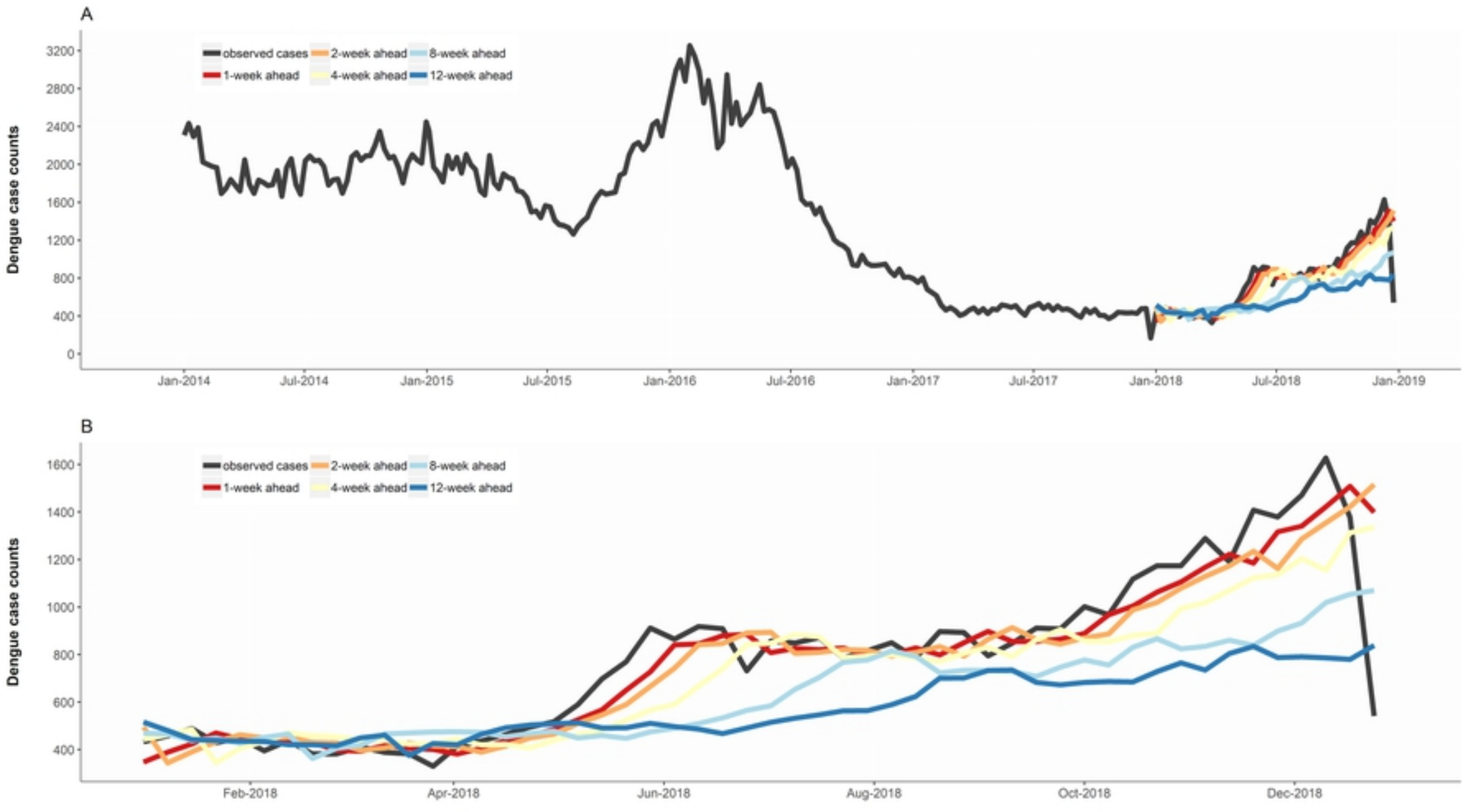
A
observed cases
2-week ahead
8-week ahead
1-week ahead
4-week ahead
12-week ahead
Dengue case counts
3200
2800
2400
2000
1600
1200
800
400
0
Jan-2014
Jul-2014
Jan-2015
Jul-2015
Jan-2016
Jul-2016
Jan-2017
Jul-2017
Jan-2018
Jul-2018
Jan-2019
B
observed cases
2-week ahead
8-week ahead
1-week ahead
4-week ahead
12-week ahead
Dengue case counts
1600
1400
1200
1000
800
600
400
Feb-2018
Apr-2018
Jun-2018
Aug-2018
Oct-2018
Dec-2018

Figure 1