ORIGINAL ARTICLE
IN ENVIRONMENTAL PUBLIC HEALTH

An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions

Sougata Deb[1], Cleta Milagros Libre Acebedo[1],
Gomathypriya Dhanapal[1], Chua Matthew Chin Heng[2]

Affiliations:

[1] M. Tech students, Institute of Systems Science, National University of Singapore, Singapore
[2] Lecturer, Institute of Systems Science, National University of Singapore, Singapore

Corresponding author:

Sougata Deb, Institute of Systems Science, National University of Singapore, Singapore.
Email: deb.sougata@gmail.com. ORCID: 0000-0001-8480-7714.

Abstract

Introduction: Dengue fever has been one of the most concerning endemic diseases of recent times. Every year, 50-100 million people get infected by the dengue virus across the world. Historically, it has been most prevalent in Southeast Asia and the Pacific Islands. In recent years, frequent dengue epidemics have started occurring in Latin America as well. This study focused on assessing the impact of different short and long-term lagged climatic predictors on dengue cases. Additionally, it assessed the impact of building an ensemble model using multiple time series and regression models, in improving prediction accuracy.
Materials and Methods: Experimental data were based on two Latin American cities, viz. San Juan (Puerto Rico) and Iquitos (Peru). Due to weather and geographic differences, San Juan recorded higher dengue incidences than Iquitos. Using lagged cross-correlations, this study confirmed the impact of temperature and vegetation on the number of dengue cases for both cities, though in varied degrees and time lags. An ensemble of multiple predictive models using an elaborate set of derived predictors was built and validated.
Results: The proposed ensemble prediction achieved a mean absolute error of 21.55, 4.26 points lower than the 25.81 obtained by a standard negative binomial model. Changes in climatic conditions and urbanization were found to be strong predictors as established empirically in other researches. Some of the predictors were new and informative, which have not been explored in any other relevant studies yet.
Discussion and Conclusions: Two original contributions were made in this research. Firstly, a focused and extensive feature engineering aligned with the mosquito lifecycle. Secondly, a novel covariate pattern-matching based prediction approach using past time series trend of the predictor variables. Increased accuracy of the proposed model over the benchmark model proved the appropriateness of the analytical approach for similar epidemic prediction research.

KEY WORDS: Climate; Covariate Pattern Matching; Dengue; Ensemble Prediction; Multiple Linear Regression; Statistics.

Riassunto

Introduzione: La febbre Dengue è stata una delle malattie infettive a carattere endemico più preoccupanti degli ultimi tempi. Ogni anno 50-100 milioni di persone vengono infettate dal virus Dengue in tutto il mondo. Dal punto di vista storico, è stata la malattia infettiva più diffusa nel Sudest Asiatico e nelle Isole del Pacifico. In tempi recenti si sono verificate frequenti epidemie di Dengue in America Latina. Questo studio si è focalizzato sulla valutazione dell'impatto di differenti predittori climatici a breve e lungo termine sui casi di Dengue. Lo studio, inoltre, ha valutato l'impatto di un modello complesso costituito da multiple serie temporali e modelli di regressione per migliorare l'accuratezza predittiva di questa patologia.
Materiali e Metodi: Dati sperimentali sono stati ottenuti da due città dell'America Latina, San Juan in Portorico ed Iquitos in Perù. Per le differenze climatiche e geografiche, San Juan ha registrato un'incidenza più alta di Dengue rispetto ad Iquitos. Usando correlazioni crociate differite, questo studio ha confermato l'impatto della temperatura e della vegetazione sul numero di casi di Dengue per entrambe le città, sebbene con gradi e gap temporali differenti. E' stata costruita e validata una strategia complessa fatta di modelli predittivi multipli attraverso un set elaborato di predittori derivati.
Risultati: La strategia predittiva complessa proposta ha ottenuto una media di errore assoluto pari a 21,55, rappresentando 4,26 punti in meno dei 25,81 ottenuti attraverso il modello standard negativo binomiale. I cambiamenti nelle condizioni cimatiche e nell'urbanizzazione sono risultati essere dei forti predittori come empiricamente evidenziato da altre ricerche scientifiche. Alcuni predittori sono risultati essere nuovi ed utili, non ancora esplorati in precedenti e rilevanti studi.
Discussione e Conclusioni: Due risultati originali sono stati ottenuti in questa ricerca. Innanzitutto, un aspetto ingegneristico focalizzato ed allineato con il ciclo vitale della zanzara vettore. Secondariamente, un nuovo approccio predittivo basato sulla corrispondenza di modelli covariati usando pregresse serie temporali degli andamenti delle variabili predittive. Un incrementata accuratezza del modello proposto rispetto al modello standard ha provato l'appropriatezza dell'approccio analitico nell'ambito della ricerca scientifica sulle previsioni riguardanti simili epidemie.

TAKE-HOME MESSAGE

Climatic conditions and urbanization have considerable impact on Aedes mosquitoes' lifecycle which subsequently affects the spread of dengue virus. Focused feature engineering can reveal these lagged relationships to form informative predictors. Additionally, ensemble prediction by combining outputs from different models is found to improve accuracy over the candidate models.

Competing interests - none declared.

Cite this article as: Deb S, Acebedo CML, Dhanapal G, Heng CMC. An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions. J Health Soc Sci. 2017;2(3):257-272

DOI 10.19204/2017/nnsm3

Received: 13/08/2017 Accepted: 02/09/2017 Published: 15/11/2017

INTRODUCTION

Dengue fever and dengue haemorrhagic fever are the predominant arthropod-borne viral diseases in the world [1]. Humans get infected when bitten by Aedes mosquitoes, the vector (carrier) of dengue virus (DENV) [2]. This paper focuses on the analysis of the occurrence of dengue fever in two cities, namely San Juan, Puerto Rico and Iquitos, Peru. Dengue outbreak is more prominent in densely populated areas, as the frequency of Aedes mosquito bites and adaptation of the vector mosquito are higher [3]. San Juan has a population density of 3,190 per km^2, whereas Iquitos has 1,222 per km^2 [4]. This makes transmission of dengue easier in San Juan than in Iquitos. In the absence of any effective vaccine, the most effective way to prevent and curb dengue transmission is by reducing the Aedes vector [5].

The ecology of DENV is inherently tied to the mosquito life cycle. The metamorphosis from egg to adult stage takes about one-and-a-half to three weeks, while the adult life span ranges from two weeks to a month depending on environmental conditions [6]. The distribution of Aedes mosquitoes is spatially and temporally dynamic, as their life cycles are short and strongly influenced by environmental factors [7]. From the Aedes mosquito bite, dengue symptoms usually start anywhere from 4 to 10 days [8]. Cooler temperatures during the early stage of the mosquito breeding cycle indicate a reduced transmission of the dengue virus [9].

Modelling of such complex relationships and interactions between diseases and climatic precursors has been recognized as a difficult problem in many studies [2, 10–12]. Gonzalez et al. used generalized additive models to capture the non-linear relationships with different weather variables [10]. Sharma et al. used advanced machine learning techniques such as artificial neural networks and support vector machines for predicting malaria outbreaks, where the latter demonstrated a significantly better prediction performance [12]. Using autocorrelation at time delay of up to 3 months and generalized linear models, a strong association was shown between temperature and rainfall with dengue fever incidence [13].

A study on the influence of meteorological factors on the dengue virus incidence in San Juan showed that these factors and dengue transmission patterns varied between years, with increased number of dengue cases peaking after higher rainfall in warmer years [14]. Through wavelet analysis, dengue incidences in Iquitos were shown to have seasonal patterns with no strong relationship with the climatic variables [15]. Similar studies revealed that climatic conditions modify the relative influence of human and climatic factors on dengue transmission patterns [16, 17]. Prediction of dengue incidence in San Juan was attempted using the NASA satellite enhanced weather forecasts with unclear model accuracy due to errors in weather forecasts [18].

The above studies demonstrate that the impact of various climatic as well as socio-environmental factors on dengue have been studied extensively all over the world over last 15 years. Eventually, there have been other studies [19, 20] summarizing, comparing and connecting findings across these studies. Multiple linear regression (MLR) and time series forecasting using Auto-Regressive Integrated Moving Average (ARIMA) models were used most frequently for predicting the number of dengue cases in these papers. A few studies also experimented with other formulations and approaches such as Poisson regression [21], negative binomial regression [22] and spatiotemporal clustering [23].

However, one common limitation of these researches was that they never explored the benefits of any ensemble prediction approaches by building and combining different predictive models with the same data. Moreover, these studies have tackled the problem either as a regression or as a time series forecasting problem. A limited number of studies that did use ensembles [24–26], have rarely used any time-series based models as candidates. Only it [26] was found to use a sequential combination of wavelet analysis, genetic algorithm and support vector machines for dengue case

prediction. However, the support vector machine was used as a learner within the genetic algorithm and not as an independent model generating a separate prediction for dengue cases. Hence, it should not be considered a stacked ensemble approach in its traditional sense.

Considering the above, the focus of this study was twofold: first, to explore the different climatic variables and identify appropriate short and long-term lagged predictors that showed strong predictive power empirically; and secondly, to build a comprehensive ensemble prediction framework by combining different time series and regression based predictions. This was done to assess the applicability and superiority of such a technique in improving prediction accuracy over the individual models.

MATERIALS AND METHODS

Data

The DengAI data was downloaded from the DrivenData website as part of a competition on predicting the spread of dengue disease [27]. The data initially came from sources supporting the *Predict the Next Pandemic Initiative* [28]. Beyond dengue surveillance data, other measurements pertained to vegetation, precipitation, and temperature. According to [29], the dengue surveillance data were provided by the U.S. Centers for Disease Control and Prevention, Department of Defense's Naval Medical Research Unit 6, Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities. On the other hand, environmental and climate data were provided by the National Oceanic and Atmospheric Administration (NOAA), an agency of the U.S. Department of Commerce. More specifically, data included:

- Normalized Difference Vegetation Index (NDVI) measurements;
- Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) satellite precipitation_measurements;
- National Centers for Environmental Prediction Climate Forecast System Reanalysis measurements;
- Global Historical Climatology Network daily climate data weather station measurements.

For the competition, the training data comprised 20 continuous features, recorded weekly from 30 April 1990 to 25 June 2010, totalling 1,456 records (Table 1).

Data exploration revealed the difference in data characteristics of the two cities, with the 20 features showing different data ranges and distribution (Figure 1).

San Juan is located along the north-eastern coast of Puerto Rico and lies south of the Atlantic Ocean. The city occupies an area of 199.2 km^2, of which 75.4 km^2 (37.83%) is water [30]. Based on the data provided, the city has a tropical monsoon climate, with an average station temperature of 27.0 ˚C, ranging from 17.8 ˚C to 35.6 ˚C between 1990 and 2008. Rainfall is distributed throughout the year, with an average station precipitation of 26.8 mm, ranging from 0 to 305.9 mm.

Iquitos is the most northern Peruvian city and has an area of 368.9 km^2. It experiences an equatorial climate, with constant rainfall throughout the year, without a distinct dry season, but a wetter summer [31]. Station temperatures range from 14.7 to 33 ˚C, with an average daily station temperature of 27.5 ˚C. The average daily station precipitation is 62.5 mm, ranging from 0 to a high 543.3 mm.

Data Preparation

The data were analysed as part of pre-processing. Actions were taken to enable model generalisation. In making these changes, care was taken to ensure that the data ranges and distribution were not adversely impacted.

Analysis of Missing Values

Analysis of the 20 features revealed that NDVI NE had 194 missing values (about 20% of available data); with additional missing values noted for all the other features. To enable generalization, the missing values were

Table 1. Descriptive statistics of original data provided.

Feature	Min	Max	Mean	Std. Dev	Outlier (3σ)	Extreme (5σ)	Null Value
Year	1990	2010					
Week of Year	1	53					
Week Start Date	19900430	20100625					
NDVI NE	-0.41	0.51	0.14	0.14	5		194
NDVI NW	-0.46	0.45	0.13	0.12	4		52
NDVI SE	-0.02	0.54	0.20	0.07	9		22
NDVI SW	-0.06	0.55	0.20	0.08	11		22
Precipitation Amt	0.00	390.60	45.76	43.72	14	3	13
Reanalysis Air Temp	294.64	302.20	298.70	1.36			10
Reanalysis Avg Temp	294.89	302.93	299.23	1.26	3		10
Reanalysis Dew Point Temp	289.64	298.45	295.25	1.53	10		10
Reanalysis Max Air Temp	297.80	314.00	303.43	3.24	3		10
Reanalysis Min Air Temp	286.90	299.90	295.72	2.57	6		10
Reanalysis Precipitation Amt	0.00	570.50	40.15	43.43	28	5	10
Reanalysis Relative Humidity	57.79	98.61	82.16	7.15	2		10
Reanalysis Saturated Precipitation Amt	0.00	390.60	45.76	43.72	14	3	13
Reanalysis Specific Humidity	11.72	20.46	16.75	1.54	3		10
Reanalysis Diurnal Temp Range	1.36	16.03	4.90	3.55	1		10
Station Avg Temp	21.40	30.80	27.19	1.29	2		43
Station Diurnal Temp Range	4.53	15.80	8.06	2.13	3		43
Station Max Temp	26.70	42.20	32.45	1.96	2		20
Station Min Temp	14.70	25.60	22.10	1.57	8		14
Station Precipitation	0.00	543.30	39.33	47.46	22	7	22
Total Dengue Cases	0.00	461.00	24.68	43.60	9	15	

imputed using multiple linear regression (stepwise selection) using the other predictor variables. This approach provided data for most missing values. For rows with missing values en-masse, the NDVI values were replaced with the immediate preceding values, while the Reanalysis and Station variables were replaced with the average of two preceding values. A block of NDVI values was missing for consecutive 14 cases in 1994. These were replaced by the average of the last two rows (progressively) and the same week of the preceding year.

Analysis of Outliers

Aside from missing values, outliers were also detected, using 3σ as inner outlier limit and 5σ as extreme limit, where σ was the observed standard deviation of the feature. Analysis of these outliers revealed that they were plausible values, and as such, they were not treated for this study.

Variable Rescaling

The wide variation in the value ranges resulted from the use of different scale, and necessitated rescaling to avoid biasing the data models. All fields were brought to comparable scales, such as °C for temperature and mm for precipitation.

Correlation Analysis

Correlation analysis of the cleaned dataset indicated that several variables had low to medium correlation with each other and with total dengue cases. The presence of variable correlation suggested possible multi-collinearity and opportunity for dimension reduction (Figure 2).

• Although correlations differed for each

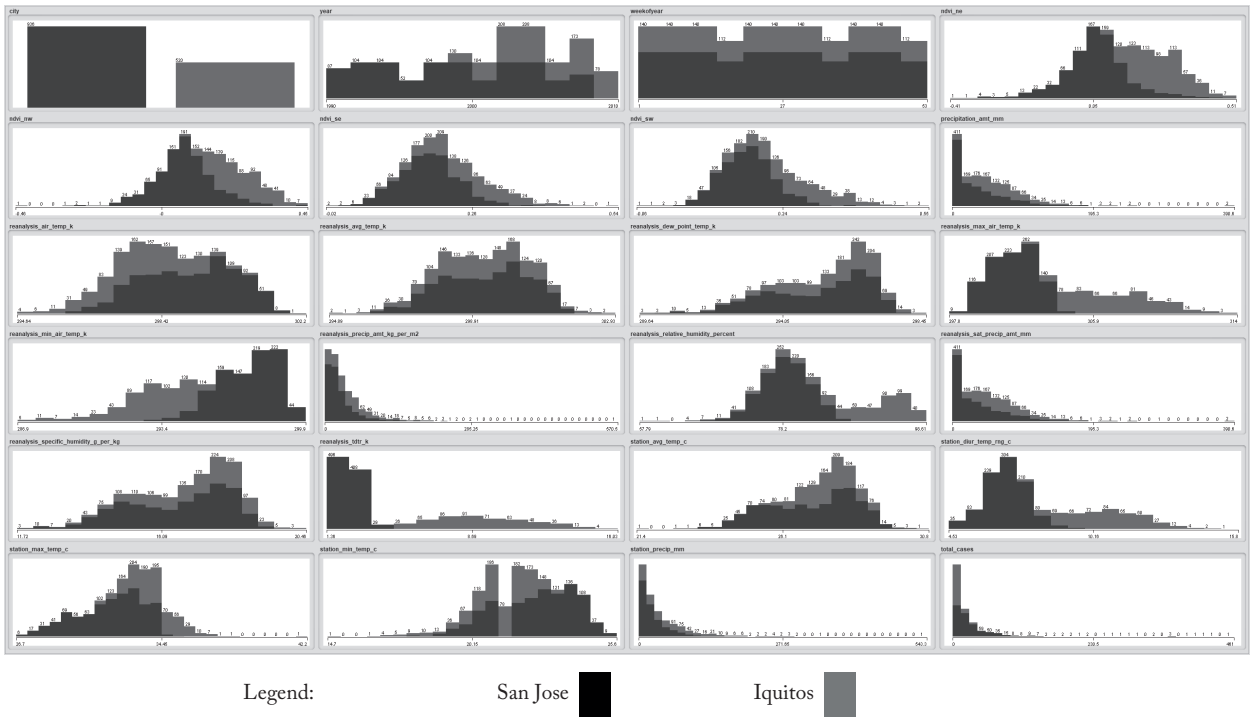

Figure 1. Data distribution by city (SJ, 936 records and IQ, 520 records).

city, the reanalysis specific humidity and reanalysis dew point temperature were the most strongly correlated with total cases. This supported the assumption that mosquitoes thrive in wet climates, which could lead to more dengue cases.

- Temperature and total dengue cases showed positive correlation, indicating higher cases of dengue during warm weather.
- In general, the precipitation measurements had weak correlation to total cases.

This presents data dimension reduction opportunity in the models.

Methods

To ensure proper rigor, objectivity and generalizability of the solutions, due importance was placed on the key modelling aspects as detailed below.

Performance Metrics

Mean Absolute Deviation (MAD) was chosen as the performance evaluation metric in line with the expectations set by the competition. MAD was calculated as:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \hat{A}_i\right| \quad (1)$$

where Ai denoted the actual dengue cases and $\hat{A}i$ denotes the predicted number of dengue cases. It should be noted here that this metric penalizes underpredictions during epidemic periods less severely than the traditional Mean Square Error (MSE) metric. This can help select the model that generates more accurate predictions for the regular / normal time periods.

Feature Engineering

Raw features available for analysis were of four different types, *viz.* temperature, precipitation, humidity and the vegetation level. Following the biological lifecycle of mosquitoes and propagation of the dengue virus, these factors were expected to impact the dengue cases at different lag periods. In this aspect, related studies have largely been different from each other and sometimes conflicting regarding the lag periods.

One study [32] found 2 months' lag for rainfall and temperature to correlate well while

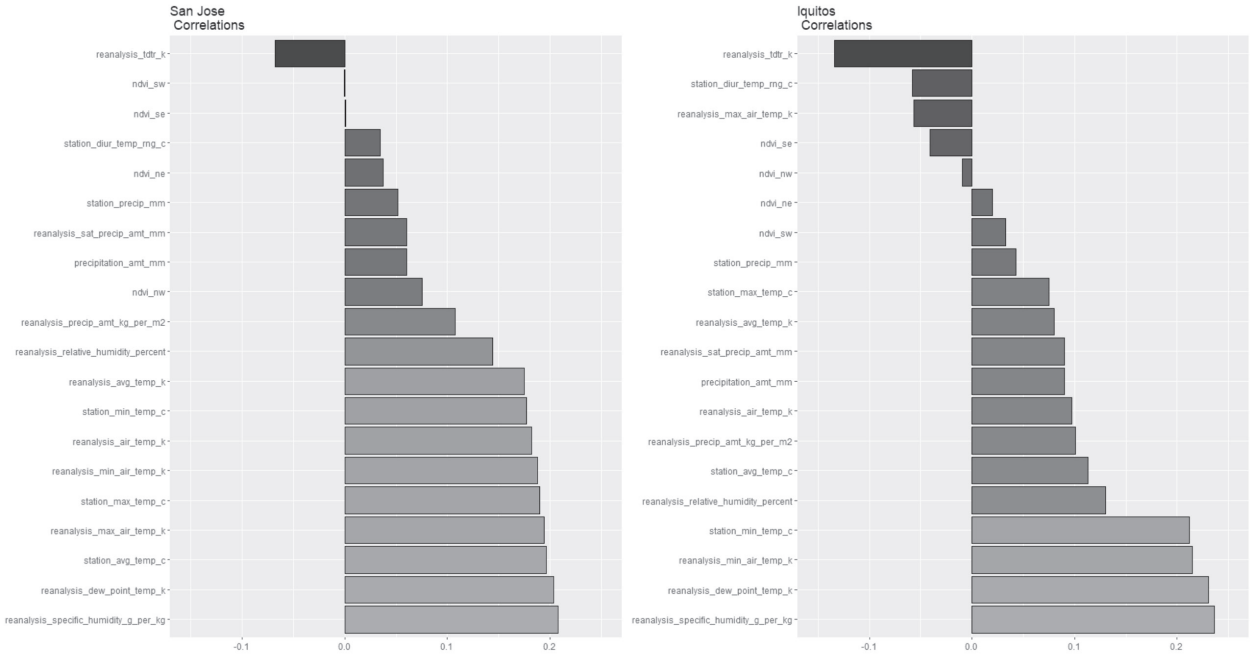

Figure 2. Correlation bar-plot for predictor importance.

another study [21] found minimum temperature of last two weeks to be a stronger predictor. Multi-wave dengue outbreak in Taiwan was found to be positively influenced by rainfall and temperature volatility because of two typhoons [33], while another study [34] found negative association with temperature with a 2 months' lag.

Based on these, it was concluded that the lagged effects of the covariates may not be uniform between the two cities under study. The same was supported by [11] which found that the impact of climatic and other predictors varied widely, driven by the geographical and tropical location of the place. Hence, the lag analysis was carried out separately for the two cities expanding up to previous 32 weeks - the maximum period that was found to explain the relationships intuitively based on mosquito lifecycle and typical propagation period of the virus.

Analytically, the appropriate lags for each variable were identified based on *lagged cross-correlation* plots using Transfer Functions. Figure 3 explains how the most suited lags were decided using the average temperature variable for San Juan as an example.

Both the most positively and negatively correlated lagged windows were retained as two different lagged predictors. Averaging over multiple weeks (*e.g.* 8-11 and 30-31 in Figure 3) helped smooth the derived variables and was expected to improve their predictive powers further.

Apart from the lagged variables, another set of derived variables were created using the cross-sectional interactions among these variables to mimic the different weather patterns known to impact the spread of dengue, either positively or negatively, *e.g.* a *hot-and-humid* variable was created by combining the temperature and relative humidity variables, a *volatile-weather* week was identified based on the difference between maximum and minimum temperature for the same week.

Finally, a decomposition based time series forecasting model was created for San Juan and Iquitos separately using the actual dengue cases. For San Juan, a *sinusoidal* seasonality was the best fit that achieved its peak during post-monsoon season. For Iquitos, a 3-point centred moving average provided the best fit for seasonality since Iquitos typically faced multiple and random monsoon sub seasons that made the method of curve-fitting infeasible. These time series forecasts were used

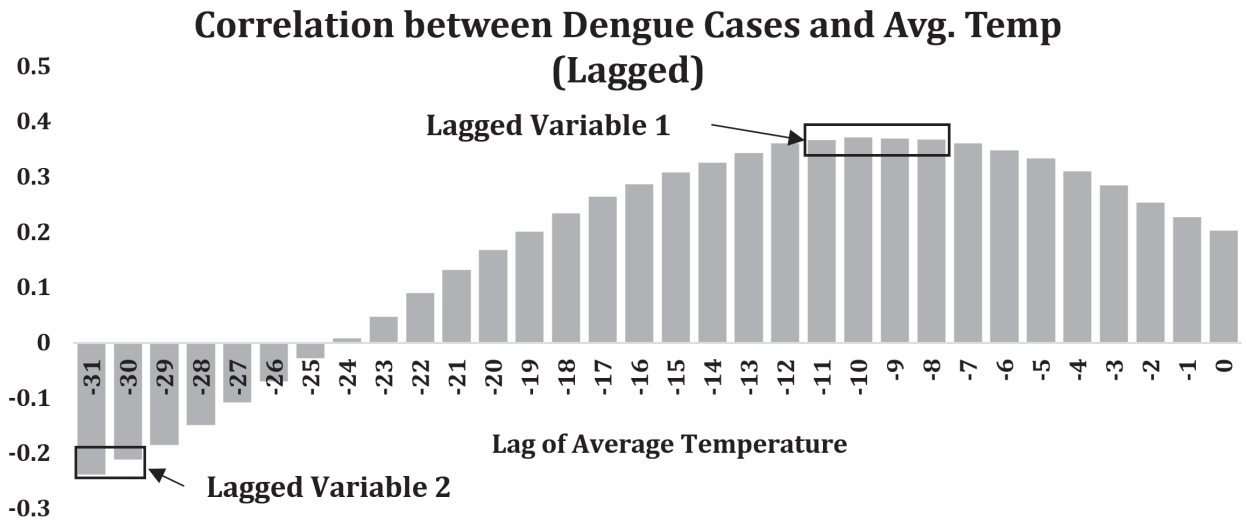

Figure 3. Lagged variable creation process based on cross-correlation.

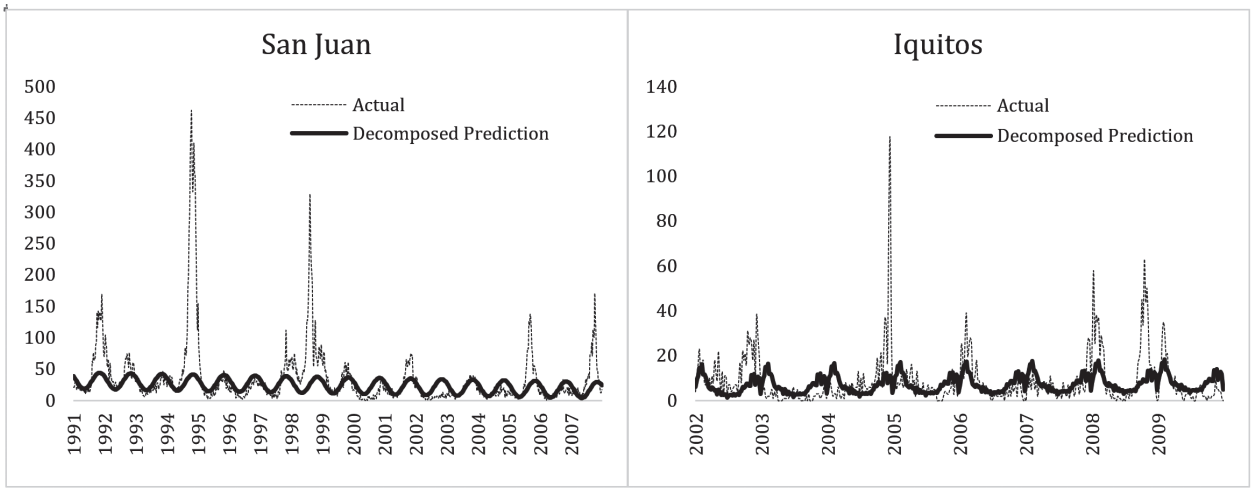

Figure 4. Time series decomposition based prediction for Dengue cases.

as additional predictors in the final models (Figure 4).

Thus, there were 103 additional predictors derived from the given features. In the reviewed literature, there are no studies that performed such elaborate and extensive feature engineering. These new derived features were expected to provide substantial lifts in the model performances.

Predictive Modelling

As discussed in the Introduction section, there have been limited attempts at combining different modelling approaches for prediction of dengue cases. Hence, an ensemble prediction framework (Figure 5) was designed and applied by using three different candidate models. The candidate models were chosen carefully to address the different aspects of the prediction goal. The subsequent sections elaborate more on these candidate models and the rationale behind their selection.

Benchmark Model: Negative Binomial Model (NGB)

Though Poisson regression was used in multiple studies, a more generalized negative binomial regression was selected to create benchmark performance on this dataset, in line with the actual competition. This was to help contrast performance of the proposed framework against an established model. The likelihood function for Negative Binomial distribution can be written as:

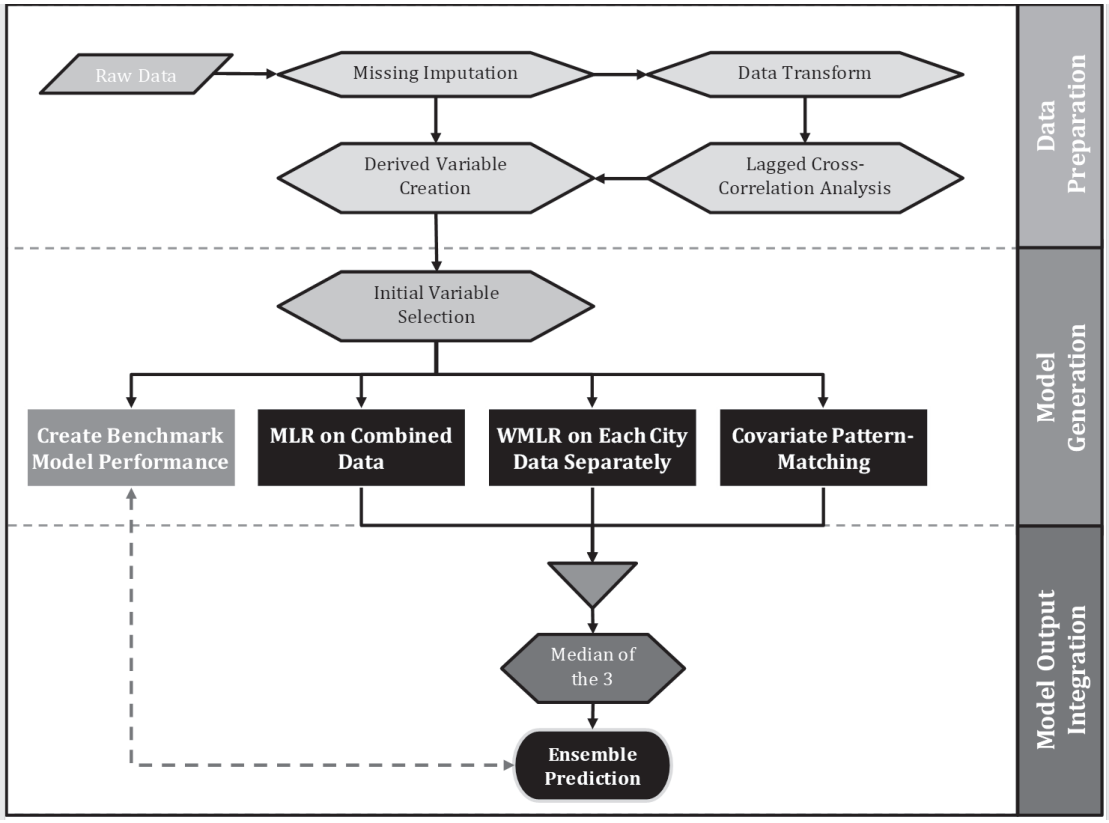

Figure 5. Model development process flow.

$$L(\beta|y,X) = \prod_{i=1}^{N} \Pr(y_i|X_i) = \prod_{i=1}^{N} \frac{\Gamma(y_i+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\mu_i+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\mu_i+\alpha^{-1}}\right)^{y_i} \quad (2)$$

where α denotes the over-dispersion parameter. As a result, Poisson regression becomes a special case of NGB when α=0. Use of NGB created a stronger benchmark performance on the dataset than a Poisson regression.

Multiple Linear Regression on Combined Data

The first model is an MLR model built on the combined data of the two cities. Though most studies recommended building different models for different geographic locations (two cities in this case), certain limitations were realized with this approach. In case a city has never faced any major dengue outbreaks in the study period (as was the case for Iquitos); the model will never learn to predict any high dengue outcomes. Building a common model will help cross-learn these different scenarios effectively to make the future predictions for both the cities more robust.

Weighted Multiple Linear Regression on Each City

To balance the possible reduction in accuracy of the first model due to its robustness, a set of city-specific regression models were built to learn the unique relationships and influences of the covariates on dengue cases. Furthermore, it was evident from Figure 4 that dengue epidemics were rare and occurred typically at 3-5 years intervals. A Weighted MLR was used by giving higher weights to these epidemic periods. A logarithmic weight as log $(100 + A_i)$ was found to provide the best results on the holdout data.

Covariate Pattern-Matching

Both the above approaches followed regression methodology and did not use any time series elements apart from the lagged covariates. Though there were several theoretical frameworks for multivariate time series forecasting and panel data analysis, they were either not applicable (*e.g.* Vector Auto-Regression) due to absence of actual dengue ca-

ses for the test period or the theoretical foundations were too complex and restrictive to get strong empirical results (*e.g.* Generalized Estimating Equations). This led us to create a simple, intuitive yet powerful methodology which delivered strong empirical results on these datasets.

There were two specific motivations behind the creation of this new methodology. Firstly, the test data had no information on the actual dengue cases which made application of any traditional time series techniques infeasible since most advanced time series models (*e.g.* ARIMA) depend on recent actuals. Secondly, it was noted that dengue cases followed certain trends of other covariates with a lag, *e.g.* dengue cases in San Juan generally increased post-summer, which is characterised by the lagged series of temperature attaining a peak at 4-6 weeks back followed by a decreasing trend in the last 4-6 weeks. Since the lags were not always fixed, it was impossible to capture these trend patterns in the individual lagged covariates that were created.

C*ovariate pattern matching* was built on the idea that a similar past pattern of covariates (C) is indicative of the likely changes in the dependent variable (A), *i.e.* the number of dengue cases. *Similarity* was defined using Euclidean distance between the latest (C_{t-k+1}, ..., C_t) and a past pattern ($C_{t-m-k+1}$, ..., C_{t-m}) based on a fixed window of length k. This memory-based algorithm then iterated over the training data range (m = 1, 2, ...) to identify the closest matching window ($C_{t-M-k+1}$, ..., C_{t-M}).

Once this lagged time point (M) was identified, change in dengue cases, expressed as:

$$\Delta A_{t-M} = A_{t-M} - A_{t-M-1} \qquad (3)$$

for the latest day in that window (t-M) was calculated. This became a prediction of the expected change in dengue cases for the latest day. The following diagram explains this pattern matching idea further using a sliding window of 26 weeks (6 months) that was found to work the best empirically (Figure 6).

The same matching process was then repeated for all 19 covariates to get 19 different predictions about the expected change over previous day's actual. Hence, the final prediction for dengue cases on day t became:

$$\hat{A}_t = A_{t-1} + \frac{1}{19}\sum_{i=1}^{19} \Delta A_{t-M_i} \qquad (4)$$

Since, recent actuals (A_{t-1}) were not available in the test data, these were replaced with the predicted values progressively to extend the time series. This approach can also be thought of as a random subspace based *k nearest neighbour* method where k=1 was considered for each subspace of predictors and the identified neighbours were subsequently aggregated by averaging.

RESULTS

The following table summarizes the covariates and their corresponding lags that were selected in each of the regression models. A red coloured lag indicates negative relationship while a green indicates a positive one. It can be observed that both small and large lags featured for different variables as significant predictors (Table 2).

Table 3 shows a summary of the individual model and ensemble accuracies, based on both the training and test data withheld for the live competition.

Predictions for the two cities using each candidate model and ensemble are shown below (Figure 7).

The output showed that the models captured the seasonal patterns in dengue cases for the two cities well. Furthermore, the *Covariate Pattern Matching* predictions on training data showed significant improvement over regression based methodologies. The results were even better than an Auto-Regressive (AR1) time series model which provided an MAD of 8.15. It should also be noted here that the performance of this model showed deterioration in the test data. This was due to the replacement of the actuals for previous days with the predicted values, since actual values were not available for test period. Using actual past counts is expected to result in a performance similar to the training data even for the test period.

DISCUSSION AND CONCLUSIONS

Multiple models, both time series and regression based, were built in this study. To allow for an objective performance benchmarking, a theoretically justified negative binomial model, with a mean absolute deviation of 25.81 based on test data, was chosen. Both MLR-based models achieved superior performance of 23.68 and 24.00 MAD respectively. This was attributed to the elaborate and extensive feature engineering which helped create meaningful derived variables from the raw covariates. Each feature selected in the final models made intuitive sense and was in-line with the prevailing weather patterns in each city.

Vegetation had a strong negative relationship, which indicated that less urbanized areas are less prone to dengue propagation. This matched the findings from other researches where unplanned urbanization has been linked to dengue spread in different geographies [20]. Similarly, high temperature and humidity in recent past were found to correlate positively with dengue spread. This too was logical, as Aedes mosquitoes are known to flourish in hot and humid weather [11]. Specifically, for Iquitos, however, larger lag in temperature (23-24 weeks) had a negative relationship. This was probably because Iquitos faces a higher average temperature throughout the year and a further increase above 30 °C during summer. Since more than 30 °C temperature becomes somewhat hostile for larva and pupa growth of Aedes mosquitoes [16], it might be working as a natural control for the subsequent weeks.

Heavy rainfall was found to have negative impact on dengue spread, again in line with other published researches [20]. Hence, this research broadly reaffirmed that climatic factors indeed impact dengue cases and with varied lags. However, most of the other researchers used lags up to 2 months (8 weeks) whereas the lags in this study spanned up to 6 months. Thus, some of the findings were new (*e.g.* negative impact of 23-24 weeks' prior temperature) and could not be validated against any published results. These new features should be studied for other locations to assess their incremental impact on prediction performances.

This study has indicated that climate change has a positive correlation with the incidence of dengue fever. This is supported by the World Health Organization (WHO) claims that changes in infectious disease transmission patterns are a likely major consequence of climate change [35]. Therefore, there is a need to study the possible underlying contributing factors and their relationships through the development of complex integrated models, in order to predict health outcomes and take the necessary preventive actions. This can help achieve one of the Sustainable Development Goals by WHO to tackle infectious diseases that are brought about by human-induced climate changes [36, 37].

Finally, the proposed *Covariate Pattern-Matching* methodology showed promising results in capturing the short-term (weekly) changes in dengue cases. Training period performance for this approach was stronger than a traditional AR (1) model. The same methodology is readily generalizable to any time series based prediction problem that is expected to have an auto-regressive property.

Finally, an ensemble approach was found to provide stronger results than each individual candidate model. Applying different modelling approaches to introduce diversity in model predictions seemed to be the key driver for this. Empirically, the MAD stood 4.26 points (25.81 vs. 21.55) lower than the benchmark prediction on this dataset.

Future research

There are three key areas where future research on this prediction problem should be pursued. First is to gather more information about the environment and response systems, along with the climatic variables. Socio-environmental factors [20] and policy-driven healthcare response systems [33, 38] were found to contain useful information about potential dengue spreads. This is reasonable

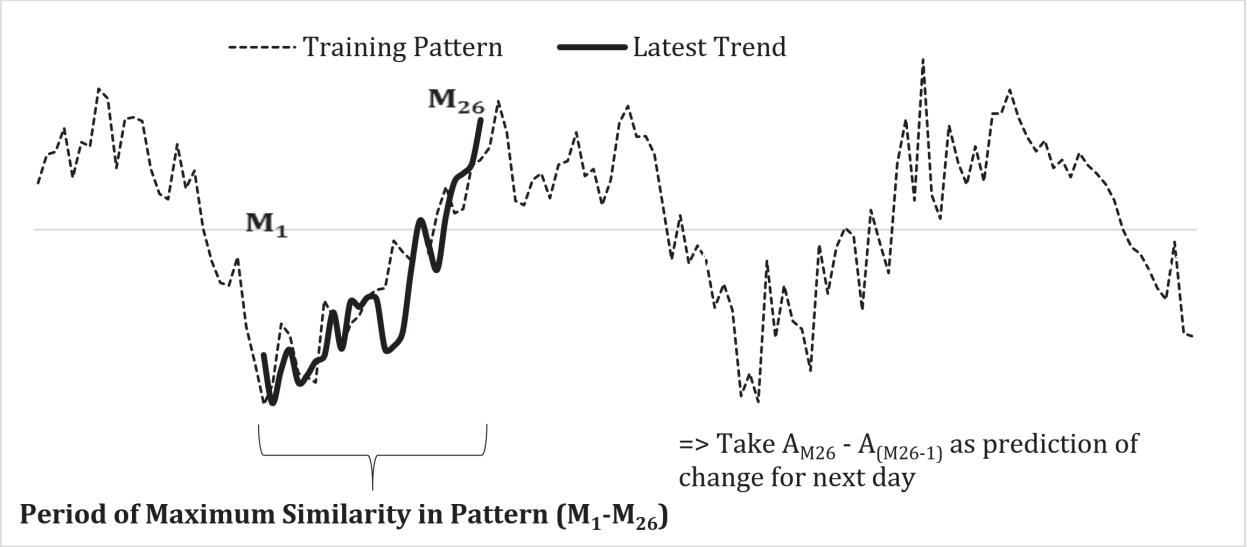

Figure 6. Demonstration of the Covariate Pattern Matching approach.

Table 2. Independent variables (numbers show lags) selected in each regression model.

Covariate	MLR-Combined	WMLR-San Juan	WMLR-Iquitos
Trend	0	0	0
Precipitation	16-23	16-23, 3-29	
Vegetation	15-21	3-20	
Max. Temp	0	0, 10-15	9-12
Avg. Temp			2-4, 23-24
Min. Temp	9-11, 18-23	30-32	15-17
Temp Range		27-32, 10-17	24-25
Rel. Humidity			15-17

Table 3. Performance details for the candidate models and final ensemble.

Model	Training Accuracy (MAD)	Test Accuracy (MAD)
Benchmark Model (NGB)	17.28	25.81
MLR on Combined Data	16.94	23.68
WMLR on City-wise Data	16.28	24.00
Covariate Pattern-Matching	7.23	26.05
Ensemble	**8.94**	**21.55**

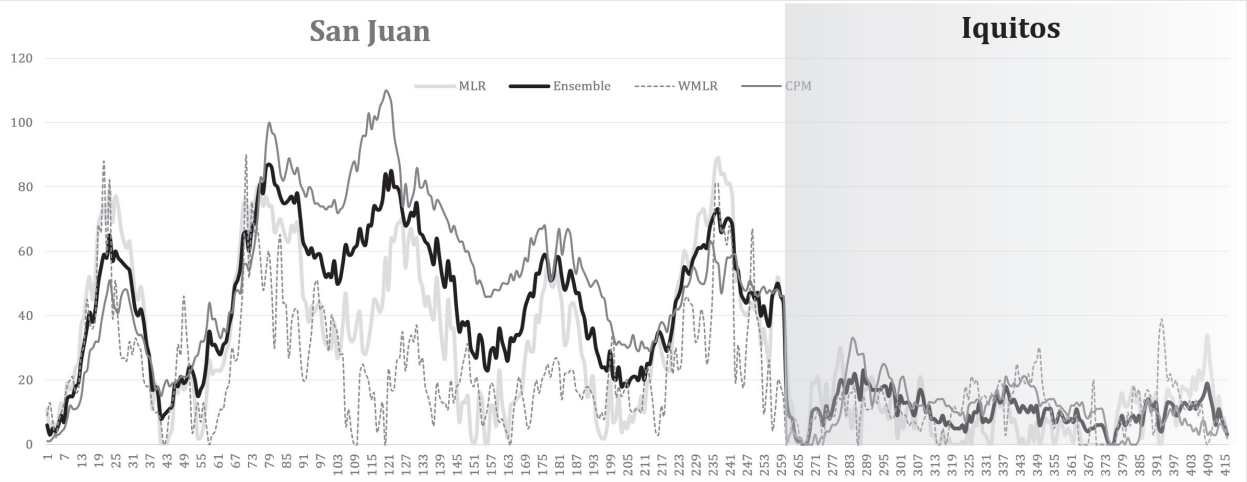

Figure 7. Weekly Dengue cases prediction for the test period.

because a locality with a strong public health control system (*e.g.* regular fumigation, waste water disposal) is expected to have lower dengue cases due to such preventive measures. Introduction of these variables can carry significant information about future outbreaks along with the climatic variables used in this study [39].

Secondly, the approaches exploited the linear relationships using different regression techniques. Application of more complex non-linear techniques such as Neural Networks, Random Forests, and Support Vector Machines can help model the latent non-linear relationships for improved performance. A caution would be to control for overfitting the training data, which was found to be a practical concern from an initial application of these techniques on this data.

Finally, the proposed *Covariate Pattern-Matching* is still a naïve approach. This methodology can be improved further, using options such as weighted averages or through variables selection based on *similarities*. Future studies should explore these and prove generalizability of this approach by applying it on other forecasting problems.

References

1. Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, et al. Dengue: a continuing global threat. Nat Rev Microbiol. 2010 Dec 1;8:S7–16.
2. Pham DN, Nellis S, Sadanand AA, Jamil J, Khoo JJ, Aziz T, et al. A Literature Review of Methods for Dengue Outbreak Prediction. in eKNOW 2016: The Eighth International Conference on Information, Process, and Knowledge Management; 2016.
3. Karl S, Halder N, Kelso JK, Ritchie SA, Milne GJ. A spatial simulation model for dengue virus infection in urban areas. BMC Infect Dis. 2014 Aug 20;14(1):447.
4. World Population Review 2017 [cited 2017 Aug 6]. Available from: http://worldpopulationreview.com/.
5. Swaminathan S, Khanna N. Viral Vaccines for Dengue: The Present and the Future. Dengue Bulletin. 2003 Dec 27.
6. Dengue Virus Net – Life Cycle of Dengue Mosquito Aedes aegypti [cited 2017 Aug 6]. Available from: http://www.denguevirusnet.com/life-cycle-of-aedes-aegypti.html.
7. Campbell LP, Luther C, Moo-Llanes D, Ramsey JM, Danis-Lozano R, Peterson AT. Climate change influences on global distributions of dengue and chikungunya virus vectors. Phil Trans R Soc B. 2015 Apr 5;370(1665):20140135.
8. Morin CW, Comrie AC, Ernst K. Climate and dengue transmission: evidence and implications. Environ Health Perspect. 2013 Nov;121(11-12):1264.
9. Alto BW, Bettinardi D. Temperature and dengue virus infection in mosquitoes: independent effects on the immature and adult stages. Am J Trop Med Hyg. 2013 Mar 6;88(3):497–505.
10. Gonzalez FC, Fezzi C, Lake IR, Hunter P. The effects of weather and climate change on dengue. PLoS Negl Trop Dis. 2013;7(11):e2503.
11. Chowell G, Cazelles B, Broutin H, Munayco CV. The influence of geographic and climate factors on the timing of dengue epidemics in Perú, 1994-2008. BMC Infect Dis. 2011 Jun 8;11(1):164.
12. Sharma V, Kumar A, Lakshmi Panat D, Karajkhede G. Malaria Outbreak Prediction Model Using Machine Learning. Int J Adv Res Comput Eng Technol. 2015 Dec 8;4(12):4415-4419.
13. Choi Y, Tang CS, McIver L, Hashizume M, Chan V, Abeyasinghe RR, et al. Effects of weather factors on dengue fever incidence and implications for interventions in Cambodia. BMC Public Health. 2016 Mar 8;16(1):241.
14. Morin CW, Monaghan AJ, Hayden MH, Barrera R, Ernst K. Meteorologically driven simulations of dengue epidemics in San Juan, PR. PLoS Negl Trop Dis. 2015 Aug 14;9(8):e0004002.

15. Stoddard ST, Wearing HJ, Reiner Jr RC, Morrison AC, Astete H, Vilcarromero S, et. al. Long-term and seasonal dynamics of dengue in Iquitos, Peru. PLoS Negl Trop Dis. 2014 Jul 17;8(7):e3003.

16. Johansson MA, Dominici F, Glass GE. Local and global effects of climate on dengue transmission in Puerto Rico. PLoS Negl Trop Dis. 2009 Feb 17;3(2):e382.

17. Morrison AC, Minnick SL, Rocha C, Forshey BM, Stoddard ST, Getis A, et. al. Epidemiology of dengue virus in Iquitos, Peru 1999 to 2005: interepidemic and epidemic patterns of transmission. PLoS Negl Trop Dis. 2010 May 4;4(5):e670.

18. Morin C, Quattrochi D, Zavodsky B, Case J. Modeled Forecasts of Dengue Fever in San Juan, PR Using NASA Satellite Enhanced Weather Forecasts [cited 2017 Aug 6]. Available from: https://ntrs.nasa.gov/search.jsp?R=20160000255.

19. 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 Infect Dis. 2014 Mar 26;14(1):167.

20. Banu S, Hu W, Hurst C, Tong S. Dengue transmission in the Asia-Pacific region: impact of climate change and socio-environmental factors. Trop Med Int Health. 2011 May 1;16(5):598–607.

21. Lu L, Lin H, Tian L, Yang W, Sun J, Liu Q. Time series analysis of dengue fever and weather in Guangzhou, China. BMC Public Health. 2009 Oct 27;9(1):395.

22. Althouse BM, Ng YY, Cummings DA. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011 Aug 2;5(8):e1258.

23. Banu S, Hu W, Guo Y, Naish S, Tong S. Dynamic spatiotemporal trends of dengue transmission in the Asia-Pacific region, 1955–2004. PLoS One. 2014 Feb 24;9(2):e89440.

24. Loshini T, Asirvadam VS, Dass SC, Gill BS. Predicting localized dengue incidences using ensemble system identification. In Computer, Control, Informatics and its Applications (IC3INA), 2015 International Conference on 2015 Oct 5 (pp. 6-11). IEEE.

25. Bakar AA, Kefli Z, Abdullah S, Sahani M. Predictive models for dengue outbreak using multiple rulebase classifiers. In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on 2011 Jul 17 (pp. 1-6). IEEE.

26. Wu Y, Lee G, Fu X, Soh H, Hung T. Mining weather information in dengue outbreak: predicting future cases based on wavelet, SVM and GA. Adv Electr Comp Sci. 2009:483–494.

27. DrivenData – DengAI: Predicting Disease Spread [cited 2017 Aug 6]. Available from: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/81/.

28. George D. Back to the Future: Using Historical Dengue Data to Predict the Next Epidemic [cited 2017 Aug 6]. Available from: https://obamawhitehouse.archives.gov/blog/2015/06/05/back-future-using-historical-dengue-data-predict-next-epidemic.

29. National Oceanic and Atmospheric Administration. Dengue Forecasting. U.S. Department of Commerce [cited 2017 Aug 6]. Available from: http://dengueforecasting.noaa.gov/.

30. Municipio Autonomo de San Juan – Informate Sanjuanero [cited 2017 Aug 6]. Available from: http://sanjuanciudadpatria.com/en/.

31. Burleigh N. Iquitos, Peru: Wet and Wild [cited 2017 Aug 6]. Available from: http://www.nytimes.com/2013/09/15/travel/iquitos-peru-wet-and-wild.html.

32. Nagao Y, Thavara U, Chitnumsup P, Tawatsin A, Chansang C, Campbell-Lendrum D. Climatic and social risk factors for Aedes infestation in rural Thailand. Trop Med Int Health. 2003 Jul 1;8(7):650–659.

33. Hsieh YH, Chen CW. Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. Trop Med Int Health. 2009 Jun 1;14(6):628–638.

34. Wu PC, Guo HR, Lung SC, Lin CY, Su HJ. Weather as an effective predictor for occurrence of dengue fever in Taiwan. Acta Trop. 2007 Jul 31;103(1):50–57.

35. World Health Organization. Climate change and human health: Risks and responses. In Climate change and human health: Risks and responses. Geneva: World Health Organization; 2003.

36. Wu X, Lu Y, Zhou S, Chen L, Xu B. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. Environ Int. 2016 Jan 31;86:14–23.

37. Chirico F. The challenges of climate change, migration and conflict in pursuit of the Sustainable Development Goals: A call to responsible and responsive policy makers. J Health Soc Sci. 2017;2(2):137–142. doi: 10.19204/2017/thch1.

38. Stone L, Olinky R, Huppert A. Seasonal dynamics of recurrent epidemics. Nature. 2007 Mar 29;446(7135):533–536.

39. Stoddard ST, Forshey BM, Morrison AC, Paz-Soldan VA, Vazquez-Prokopec GM, Astete H, et. al. House-to-house human movement drives dengue virus transmission. Proceedings of the National Academy of Sciences. 2013 Jan 15;110(3):994–999.