RESEARCH

Prediction of Dengue Outbreaks Based on Disease Surveillance, Meteorological and Socio-Economic Data

Raghvendra Jain[1*], Sra Sontisirikit[2], Sopon Iamsirithaworn[3] and Helmut Prendinger[1]

*Correspondence:
raghavendra.jain@gmail.com
[1]National Institute of Informatics, Tokyo, Japan
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: The goal of this research is to create a system that can use the available relevant information about the factors responsible for the spread of dengue and; use it to predict the occurrence of dengue within a geographical region, so that public health experts can prepare for, manage and control the epidemic. Our study presents new geospatial insights into our understanding and management of health, disease and health-care systems.

Methods: We present a machine learning-based methodology capable of providing forecast estimates of dengue prediction in each of the fifty districts of Thailand by leveraging data from multiple data sources. Using a set of prediction variables we show an increase in prediction accuracy of the model with an optimal combination of predictors which include: meteorological data, clinical data, lag variables of disease surveillance, socioeconomic data and the data encoding spatial dependence on dengue transmission. We use Generalized Additive Models (GAMs) to fit the relationships between the predictors and the clinical data of Dengue hemorrhagic fever (DHF) using the data from 2008 to 2012. Using the data from 2013 to 2015 and a comparative set of prediction models we evaluate the predictive ability of the fitted models according to RMSE and SRMSE.

Results: In this paper, we present a model which allows for combining different predictors to make forecasts and also describe the statistical significance of the variables used to characterize the forecast. The discriminating ability of the optimal model was evaluated against Bangkok specific and WHO threshold of the epidemic in terms of specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV).

Conclusions: Along with the high accuracy of prediction of our model, the novel contribution of this research is also that we determine that for the prediction of dengue outbreaks within a district, the influence of dengue incidences and socioeconomic data from the surrounding districts is statistically significant. This suggests the influence of movement patterns of people and spatial heterogeneity of human activities on the spread of the epidemic.

Keywords: Dengue forecasting; Data-driven Epidemiology; Disease Surveillance; Generalized Additive Models (GAMs)

1 Background

Dengue, a mosquito-borne viral disease, is caused by four distinct, but closely related, serotypes of the virus [3, 4]. Recovery from infection by one of these four (DEN-1, DEN-2, DEN-3, and DEN-4) provides the infected person lifelong immunity against that particu-

lar serotype and cross-immunity to the other serotypes. The time duration for this cross-immunity is 6-12 months [3]. If the person is infected by other serotypes subsequently then the risk of severe dengue increases. The uninfected mosquitoes get the virus from infected humans and thus the later becomes the primary carrier, multiplier, and transmitter of the DENV (dengue virus).

Thailand began to experience Dengue fever in 1949 and it became pandemic in the country for the first time in 1958 in Bangkok [5]. The information about the clinically diagnosed cases of dengue fever(DF), dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) are sent to the Epidemiology Department in Bangkok [6] via the Provincial Health Offices. The transmission of DENV which occurs through the bite of infected Aedes mosquitoes, principally Aedes aegypti, has dramatically increased in the last two decades [7] and occurrence of dengue fever is likely to rise [8].

A combination of several factors determine the risk of DENV transmission. These factors include density of infected mosquitoes, (level of) immunity of people to dengue serotypes, meteorology, human related factors e.g. housing type, population density and demographics, cleanliness etc. Several potential predictive indicators that contribute to the above mentioned factors have been described[9–12]. For example, the density of infected mosquitoes depends on total mosquito density [13, 14] which depends on inside house breeding [15] and outside house breeding [12]. The inside house breeding depends on cleaning activity [16, 17] and housing type [18] among other factors whereas outside house breeding is primarily influenced by neglected garbage [19], population density [20–22] and rain [23]. The seasonal influence on rain [24, 25] and temperature [25–27] contributes to the mosquito density [11, 12, 23, 26, 28]. The density of infected mosquitoes in the region is influenced by dengue rate in the region [29, 30] and so is the number of people with immunity [14, 31] along with the population demographics [8, 32, 33].

Despite the numerous urban outbreaks of dengue with significant health and economic impact [34–37], the detailed surveillance for diagnosing dengue has been limited which makes it difficult to generate detailed information on its epidemiology [38, 39]. Moreover, there are currently no licensed vaccines or specific therapeutics for the treatment of the infected people. A variety of dengue vector control strategies have been adopted in different regions [40–43] but this has not stopped its rapid emergence and global spread [44]. Therefore, the goal of this research is to create a system that can use the available relevant information about the factors responsible for the spread of dengue and; use it to predict the occurrence of dengue within a geographical region, so that public health experts can prepare for, manage and control the epidemic. Our study presents new geospatial insights into our understanding and management of health, disease and health care systems. It yields practical results (e.g., results of value to a national public health, control, screening or prevention program, or local resource planning program along with serving to re-demonstrate a previously well-documented phenomenon.

2 Methods

2.1 Study Area

The dengue outbreak in Bangkok can affect to dengue situation for the whole country because Bangkok is a very crowded city located at the center part of Thailand. In the 2010 census, the population of Bangkok was about 8.28 million, although just 5.7 million were the registered residents. The population within the city during the day swells to about to

15 million [24] due to the commuters from the surrounding areas. During a winter season, the temperature in Bangkok is still high around 28-35 degree Celsius and there is rain in every season [45]. Figures 2 and 1 show the mean monthly diurnal temperature range (DTR) and cumulative monthly rainfall of Bangkok throughout the years (2008 – 2015). The diurnal temperature range (DTR) is the difference between the daily maximum and minimum temperature. The experimental evaluation has shown that DTR influences two important parameters underlying dengue virus (DENV) transmission by Aedes mosquito [26].

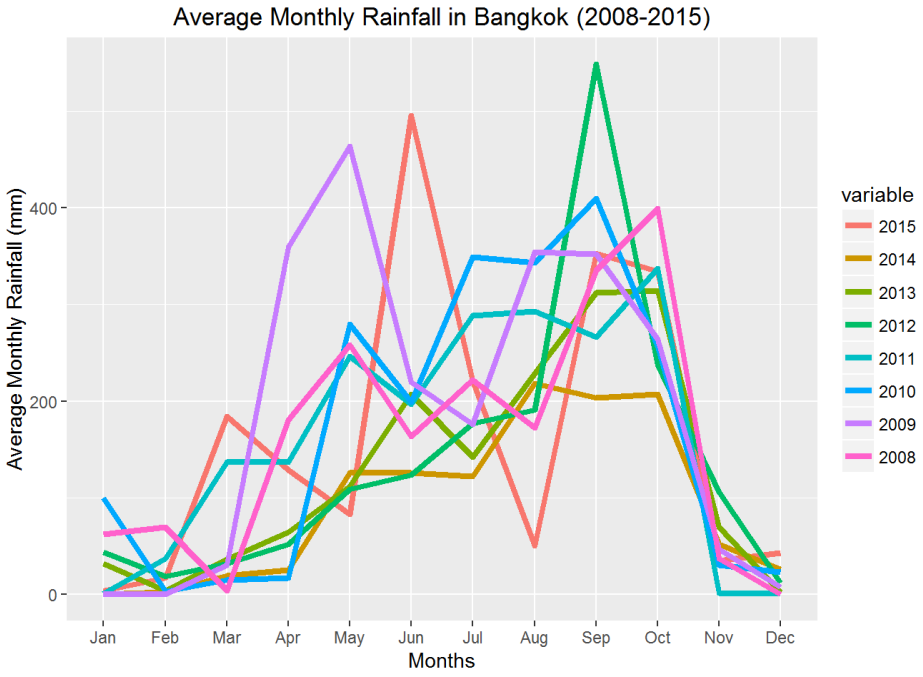


Figure 1: Rainfall within Bangkok throughout the years.

Moreover, many people who live in Bangkok travel to other provinces frequently, so these people may be the main carriers and multipliers of the virus that cause dengue outbreak in other parts of Thailand. The Bangkok city covers an area of 1,568.737 square kilometers and it is subdivided into 50 districts, which are further sub-divided into 169 sub-districts. The total population registered in Bangkok is 5,693,884 and more than three million un-registered people live in Bangkok. The average registered population of Bangkok districts in 2014 is 113845. The most populated district is *Saimai* (194,511 residents) and the least is *Samphanthawongse* (26359 residents). But the highest and lowest population density district are *Phranakhon* (157791 men/km 2) and *Thawiwatthana* (2948 men/km 2).

2.2 Data

Climate is one of the main factor related to dengue outbreak both locally and globally [46, 47]. Most of Thailand has a tropical wet and dry climate type, making the city amenable for Aedes mosquito to breed and spread in any season. The role of the variation in climatic factors on transmission dynamics and the geographic distribution of dengue has been well-studied [48]. The rainy season allows Aedes mosquitoes egg to grow into adult mosquitoes

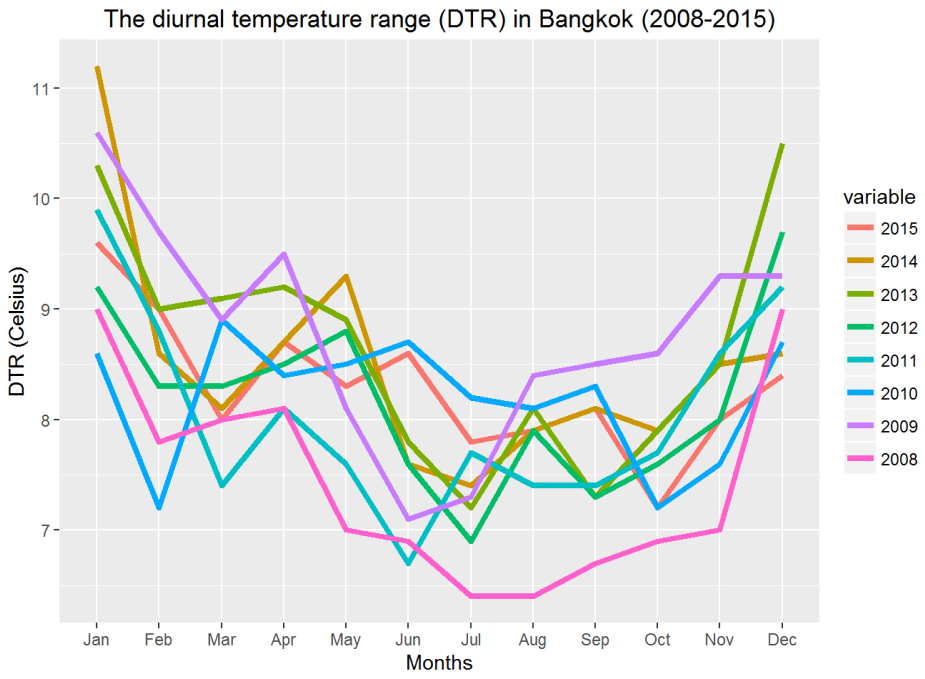


Figure 2: DTR within Bangkok throughout the years.

easier than in dry season. The increase of Aedes mosquitoes is directly affecting dengue cases in Thailand as shown in Figure 4.

Figure 3 shows the increase of DHF cases in Bangkok up to 2015. The average number of DHF cases in Bangkok from 2008 until 2015 is around 4,750 cases per year. The highest number of DHF cases in Bangkok is 16942 which happened in 2013 followed by 14154 cases in 2015. As shown in Figure 4, the highest number of monthly DHF cases in Bangkok is 11752 in the month of November followed by 9511 cases in October.

Figure 5 shows the variation of dengue cases with months. In the rainy season (mid-May–October), dengue cases rise dramatically, and then they decrease suddenly in summer season (February—mid-May). The reason that November has the highest number of cases is that dengue virus in patients need 4–10 days for the incubation period. So most dengue patients in November are infected in the rainy season (October-November).

In this study, we have used aggregated monthly dengue cases (dengue hemorrhagic fever) at district level from the Department of Disease Control, Bangkok for the period 2008 – 2015. Rainfall data from 2008 – 2015 is provided by the Department of Drainage and Sewerage, Bangkok and the temperature data for the same period is provided by the Meteorological Department, Thailand. Records of cumulative monthly rainfall (mm) and mean monthly DTR readings (°C) were obtained. The meteorological data were merged to the aggregated total number of confirmed dengue cases per month in each district of Bangkok. Using the geographical information of different districts of Bangkok, this data of each target district was merged to the aggregated total number of confirmed dengue cases per month in each of its surrounding district. Thus, our data has (a) Dengue virus prediction variables (Monthly DHF incidents in a district + aggregated monthly DHF incident in the

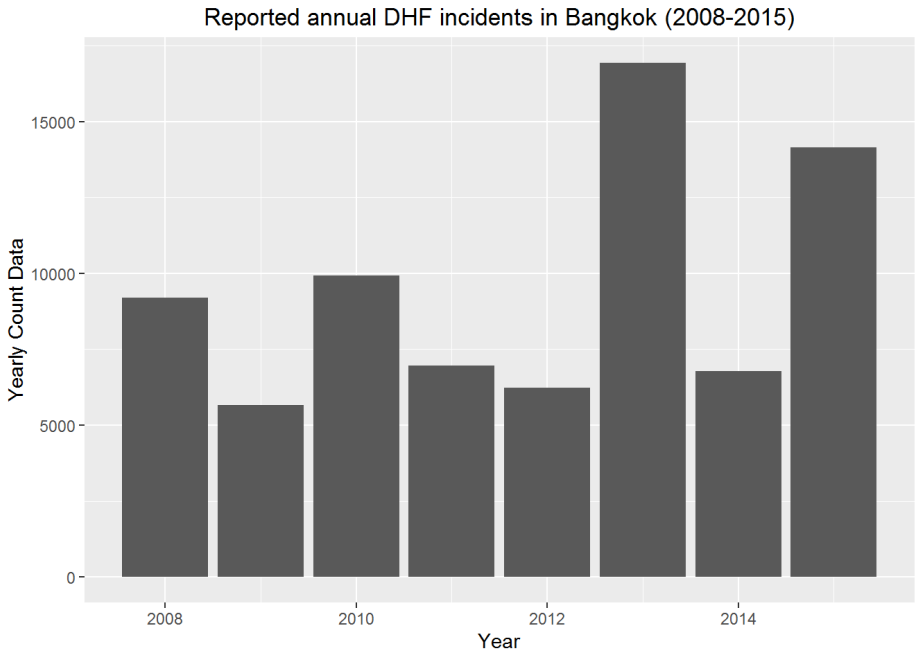


Figure 3: The DHF incidence peaked in 2013 and 2015. It shows that dengue outbreak increases every alternative year. Data from [49]

.

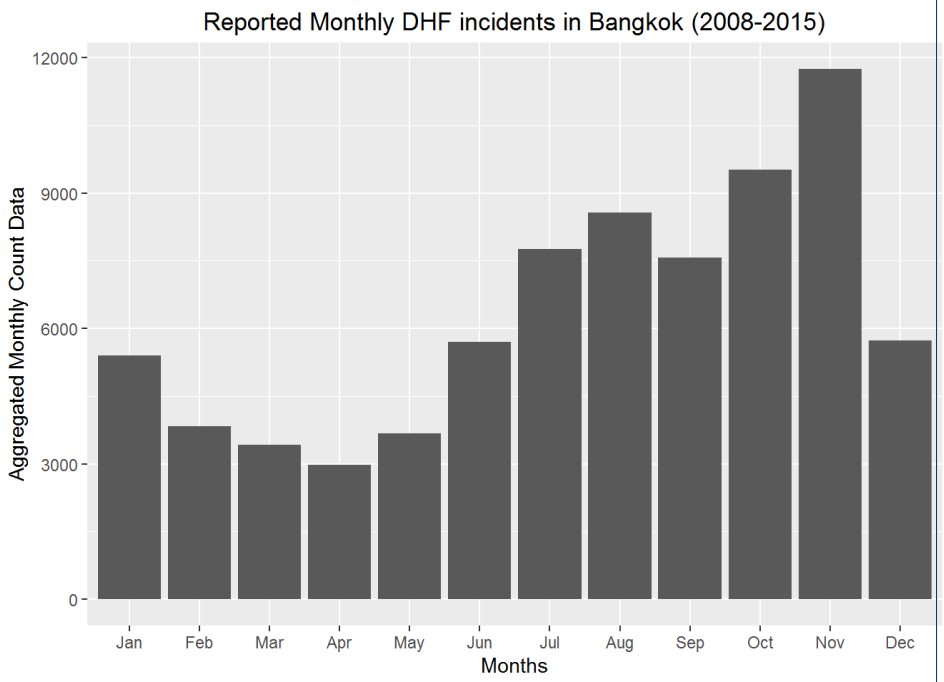


Figure 4: The DHF incidence peaked in October and November. Graph shows the amount of dengue patients for each month for the years 2008–2015. Rainy season continues from mid-May to mid-October. Data from [49]

.

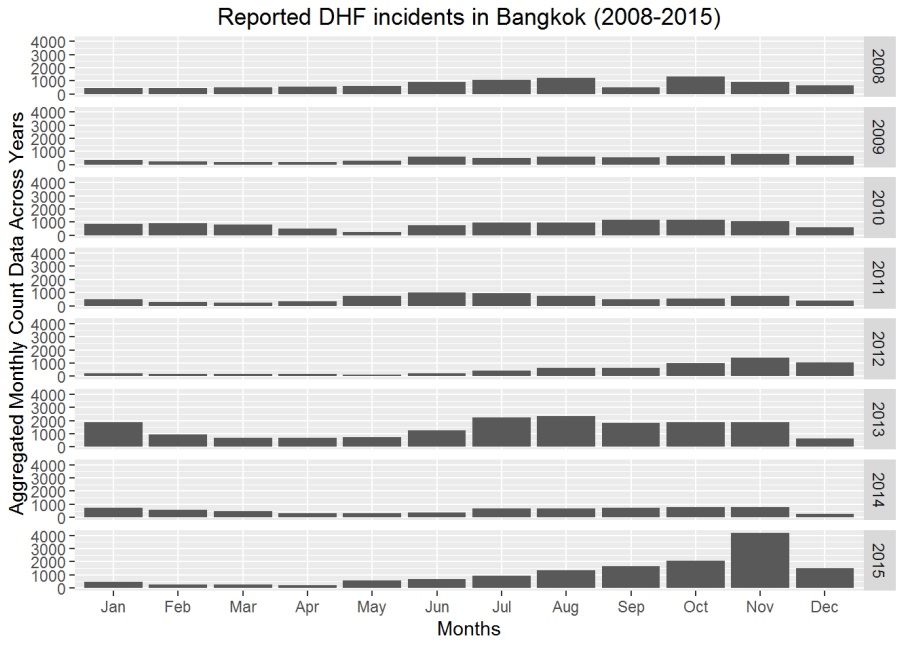


Figure 5: The DHF incidence peaks in the months of October and November. However in the year 2013, DHF outbreak happened for several months continuously.

nearby districts) and (b) Aedes mosquito prediction variables (cumulative monthly rainfall in the Bangkok + mean monthly average diurnal temperature range (DTR) in Bangkok).

A model was developed and validated by dividing the data file into two data sets: one for training the model and another for testing and validation of the fitted model. Here we varied the period in both the data sets. For example, in one case of prediction the data from January 2008 to December 2012 was used to train a model, and data from January 2013 to December 2015 was used for testing and validation of the fitted model. All the analyses were performed in R [50] using the mgcv package [51]. The details of the different such cases of predictions are listed in Table 1.

2.3 Statistical Analysis

These analyses focus on studying the following relationships:

- relationship of meteorological (DTR and rainfall) and socioeconomic data (monthly garbage collection in each district) to the time series of dengue incidences in that particular district.
- relationship of dengue transmission in a particular month in a particular district with the data of its past occurrences.
- relationship of dengue transmission in a particular month in a particular district with the data of past occurrences of dengue in its surrounding districts.

We created a set of prediction models encoding the above-mentioned relationships. Our geospatial/statistical method used in our work and the approach using Generalized Additive Models (GAM) to derive the insights are similar to research study [52] conducted in Indonesia. However, along with evaluating different predictions models, we evaluate a different/unique hypothesis using a novel data-set; the feature-set used in the study is richer and the study area/country is different from the above-mentioned work.

The target of prediction was the cumulative dengue count in Bangkok (i.e. sum total of dengue incidences in all the 50 districts of the city) in a particular month of the year. According to the previous studies to determine the optimal lead time for dengue forecasts [53], there is evidence of increasing dengue cases in lag time of up to 4–20 weeks. Thus, similar to [52] we decided a priori, for meteorological variables, to use lag times with up to 4 months delay (i.e. 0–3 months) in the analysis. Since the dengue counts vary within and between the years, the count data is likely to be over-dispersed. Thus, rather than using the "standard" Poisson regression in which it is assumed that variance of count data is constant regardless of the expected value, we adopt a Quasi-Poisson regression in which the variance of count data (dengue counts) is assumed to be a linear function of the mean. To allow for over-dispersion a log-link function of dengue count data is used. To allow for non-linear response and exposure association between the predictors and the dengue incidences, cubic splines of 3 degrees of freedom was applied on the meteorological variables. The generalized additive model (GAM) can be expressed as:

$$log(C_{0,t}) \sim \alpha + \sum_{l=0}^{3} ns(\text{T}_{lt}, d = 3) + \sum_{l=0}^{3} ns(R_{lt}, d = 3) \quad (1)$$

where log() denotes the natural logarithm, C represents the total dengue count data, t denotes to time in months, T represents DTR (°C), R denotes the mean monthly rainfall (mm) and l, denotes the lag variables, ns() denotes a natural cubic spline.

Disease surveillance data of each districts as predictor

Since the current number of dengue incidences are influenced by the number of cases in the past, to determine this period of influence we have considered two approaches. The first approach focuses on determining the optimal lag term for short-term lagged dengue incidence data. The auto-regressive patterns in dengue time series data were studied by fitting a GAM in which data up to a delay of 4 months was used (similar to what we did with the meteorological data). This model was to fit to assess the influence of past dengue incidence on current count independent of meteorological factors. The regression model can be expressed as:

$$log(C_{0,t}) \sim \alpha + \sum_{l=1}^{4} ns(C_{lt}, d = 3) \quad (2)$$

Second approach focused on assessing the risk of retrospective transmission (1–48 months) on dengue incidences. Prior studies conducted in the region have shown that cross-immunity for dengue virus serotypes may significantly alter the dengue transmission over a period [31, 54]. We hypothesize that this might partly explain bi-annual cyclic epidemic pattern of dengue occurrence in Bangkok as shown in Figure 3. Thus, we assess and estimate the risk of retrospective dengue transmission up to 1–30 months on current dengue transmission. To incorporate the effects delayed in time, the statistical model of DLNM[1] was used to describe the additional time dimension of this exposure-relationship[55].

[1]The GitHub repository of R implementation of DLNM used for our analyses is available at https://github.com/gasparrini/dlnm

The regression model can be expressed as:

$$log(C_{0,t}) \sim \alpha + \sum_{l=1}^{30} DLNM(C_{lt}, d = 4) \tag{3}$$

The output from both these approaches was then combined to determine the 'optimal lag' of disease surveillance.

Disease surveillance data from surrounding districts as predictor There is much research which concludes that many dengue cases that occur in urban areas are due to the factors such as high population density, inadequate housing, and inappropriate human behavioral practices [20–22]. Surveillance of Aedes mosquito density is important for construction models of dengue transmission, in order to prioritize areas and seasons for vector control. The 80% of larvae or pupa in house are from Aedes mosquito. A recent study [12] has explored the dengue occurrence in a region in relation to its surrounding regions. The study is conducted in near real-time using object-based and spatial metric approaches. The geospatial analysis conducted on the data acquired using Google search and advanced land observation satellite images suggests that the occurrence and spread of dengue cases are positively correlated with densely populated areas which are *surrounded by dense vegetation*. This further suggests that the spatial heterogeneity of human activities influence the dengue epidemic. Thus, to determine the influence of spatial heterogeneity of human activities ongoing in nearby areas, we consider the data (both 'short-term' and 'long-term') from the dengue incidences of surrounding districts. The regression model can be expressed as:

$$log(C_{0,t}) \sim \alpha + \sum_{l=1}^{4} ns(S_{lt}, d = 3) + \sum_{l=0}^{30} DLNM(S_{lt}, d = 4) \tag{4}$$

where log() denotes the natural logarithm, C represents the total dengue count data, S represents the count data of dengue incidence occurred in the surrounding districts, t denotes to time in months, l denotes the lag variables and ns() denotes a natural cubic spline.

Waste disposal data from each district as predictor Previous studies have shown the spatial correlation of socioeconomic data and urbanization with dengue incidences [56, 57]. Since the waste disposal and landfill dumps are the spatial infrastructures of any modern city, we used the data about monthly garbage collection from each district as an indicator for social capital.

Thus using the above-mentioned predictors, we evaluated the following models:

(A) Optimal Meteorology Model (includes lag of 0-4 months for DTR and mean monthly rainfall)
(B) Short-term Surveillance Model (may include lag of $1-4$ months)
(C) Optimal Lag Model (may include lag of $1-30$ months)
(D) Combination of (A) and (C)
(E) Combination of (D) with lagged dengue incidences of surrounding districts
(F) Combination of (E) with garbage collection data of each district as the social capital

3 Results

Using the lags of month 0–3 of Diurnal Temperature Range (DTR) and monthly average rainfall data in the model, we analyzed the (1) statistical significance of the model terms (2) association of DTR and rainfall with dengue cases.

As shown in Table 4 the statistical effects appear at DTR and rainfall for all the lags. As shown in Figure 6, the association of temperatures with dengue cases in lag 0,1 decreases when the temperature is less than 8 °C and then increases. However, for the lag 3, this association slightly increases and then decreases. For lag 4, it is observed that the association of temperature with dengue cases increases linearly when the temperature is less than 9 °C and then a strop drop is observed. Due to the statistical significance of these terms i.e. lags of month 0–3 for the meteorological data, we call the model comprising of them as 'Optimal Met' model (model A). The performance of the model using the training data from $2008-2012$ calculated using R-squared (adjusted) is 0.283, RMSE is 8.462 and SRMSE is 0.52. The 'Deviance Explained' is 31.9%. The visual analyses of the prediction of the 'Optimal Met' model are shown in Figure 11a.

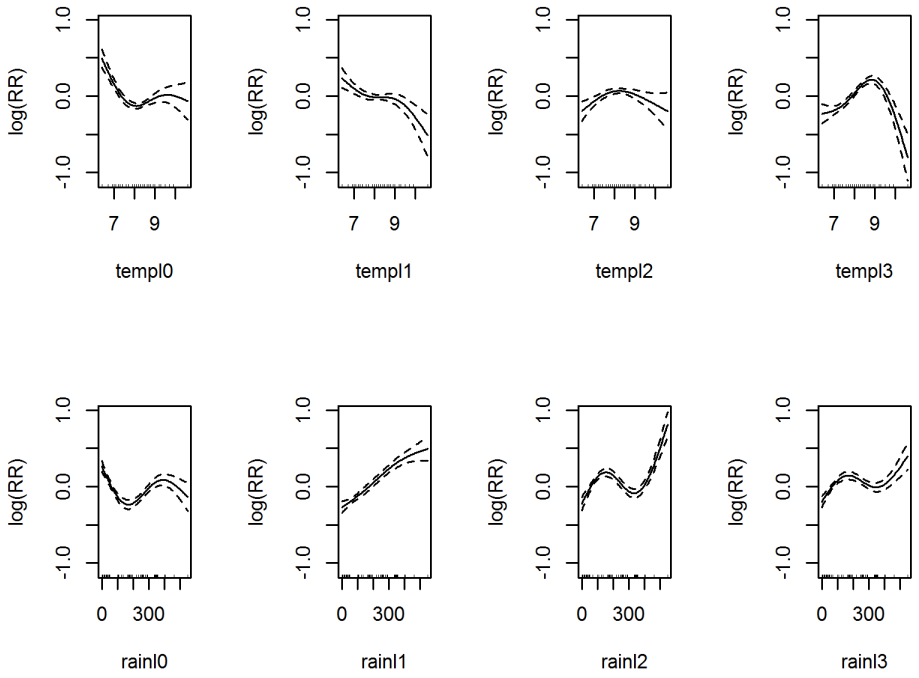


Figure 6: **Association between the meteorological variables and dengue over lags of 0–3 months**. Solid lines represent relative risks (RR) of dengue cases whereas the dotted lines depict the upper and lower limits of 95% confidence intervals. P-values and the initial analyses are listed in Table 4

.

We calculated the cross-correlation of dengue cases with (a) dengue cases of lag of 0–3 months (b) DTR of lag of 0–3 months (c) rainfall of lag of 0–3 months and (d) the dengue cases of lag of 0–3 months from the surrounding districts. As shown in Figure 15 the

cross-correlation indicate that the highest positive association between dengue incidence and lagged dengue incidences were found at lag 0 (r, 0.667), with rainfall at lag 2 (r, 0.428) and with dengue incidences from surrounding districts at lag 1 (r, 0.514). There was a negative correlation of DTR with dengue incidence at all the lags. For lag 0 (r, -0.261), lag 1 (r, -0.373), lag 2 (r, -0.309) and lag 3 (r, -0.154) was observed.

As shown in Figure 7 the analyses performed using the data of lagged dengue incidences, when lag terms of 1-4 months were included, show that dengue transmission increases almost linearly with the lagged dengue incidences at lag 1 and lag 2 till the dengue counts is less than 30. Afterwards as well as non-linear increasing trend is observed. While, dengue cases linearly decrease with lag 3 and lag 4 and these terms were not statistical significant ($p > 0.05$). Thus, for a model in which lag-terms of up to 2 months were included is termed as 'Short-term Lag Surveillance' model (model B). The performance of the model using the training data from $2008 - 2012$ calculated using R-squared (adjusted) is 0.4, RMSE is 7.9 and SRMSE is 0.49. The 'Deviance Explained' is 41.0%. The visual analyses of the prediction of the 'Short-term Lag Surveillance' model is shown in Figure 11b.

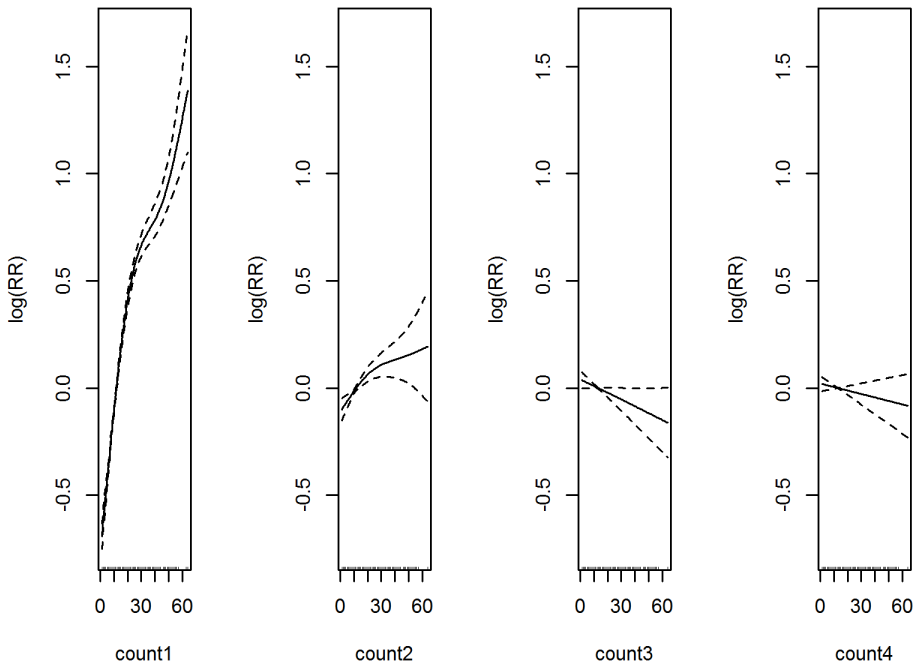


Figure 7: **Short-term Lag Model**: Association between past dengue count over lags of 1-4 months and the dengue outbreak. Solid lines depict relative risks (RR) of dengue cases while the upper and lower limits of 95% confidence intervals are shown by the dotted lines.

Meanwhile, when the long-term lagged dengue incidence data was taken into account, the non-linear distributed lag models were used using dlnm package [58]. The influence of long-term lagged data of dengue incidences is shown using the contour plot in Figure

8a. It is observed that lagged long-term data has lower relative risks of transmission up to almost 2 years following a large outbreak in around lag 24. This suggests a negative feedback cyclic pattern. Figure 8b suggests that when an outbreak happens in a particular month then dengue risk in each of the following months will increase with a peak in next 23 months (Figure 8b). Both Figure 8b and Figure 16 are the lag-response curve for the differing number of dengue cases after an outbreak happens in a specified month.

Thus based on these analyses, the optimal variables for the prediction models included dengue count at lag 1,2 for short-term lags (as used in 'Short-term Lag Surveillance' model) and at lag 23. The model that consists of these lags is termed as 'Optimal Lag Surveillance' model (model C). The performance of the model using the training data from $2008-2012$ calculated using R-squared (adjusted) is 0.53, RMSE is 7.07 and SRMSE is 0.43. The 'Deviance Explained' is 53.0%. The visual analyses of the prediction of the 'Optimal Lag Surveillance' model is shown in Figure 11c.

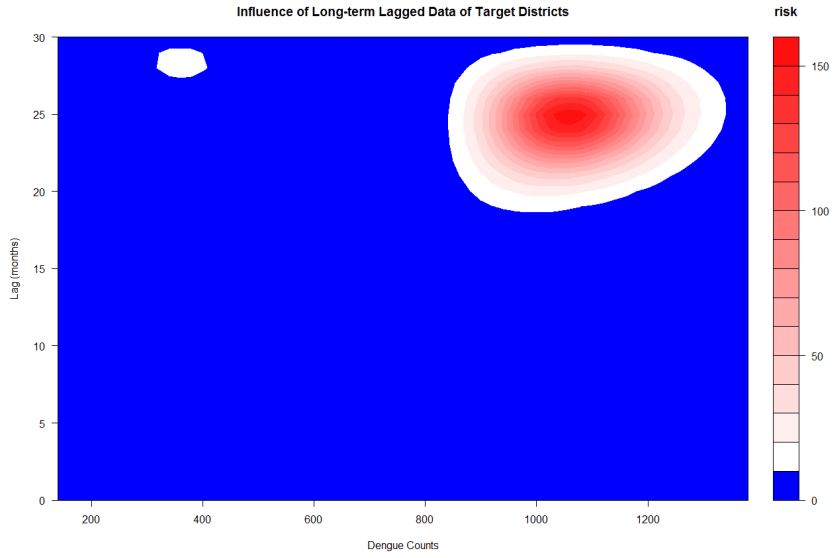
We then combine the meteorological variables used in 'Optimal Met' model (model A) and dengue lag terms used in 'Optimal Lag Surveillance' model (model C) to determine their association with dengue incidences. This model is termed as 'Optimal Met and Lag Surveillance Model' (model D). The performance of the model using the training data from $2008-2012$ calculated using R-squared (adjusted) is 0.69, RMSE is 5.72 and SRMSE is 0.35. The 'Deviance Explained' is 70.0%. The visual analyses of the prediction of the 'Optimal Met and Lag Surveillance Model' model is shown in Figure 11d.

Since one of our hypotheses is that the significance of movement patterns of people and spatial heterogeneity of human activities on the spread of the epidemic in statistically significant. In other words, the dengue cases in a particular district are influenced by the dengue cases in there surrounding districts. To test the hypothesis, we determine how the occurrence of dengue in a target district is influenced by the occurrence of dengue in its surrounding districts. Both short-term and long-term lagged dengue cases data of the dengue incidences in surrounding districts was taken into account. For short-term lags, we considered the lagged data of past 1-4 months in which the data of lag 1 and 2 months was found to be statistically significant ($p < 0.05$). For long-term lags, the data up to the past 30 months was used. The non-linear distributed lag models were used using dlnm package [58] to determine the relative risks of transmission following an outbreak in a specified month.

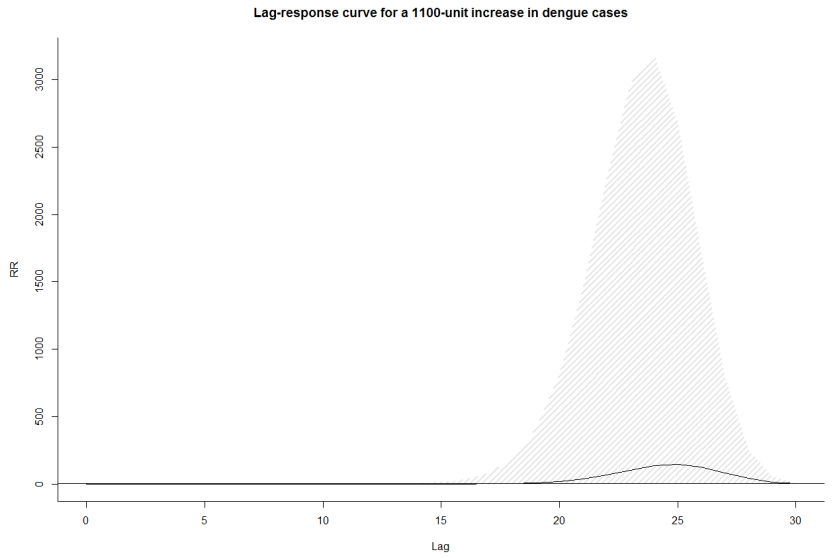
Figure 9 suggests that when a large outbreak happens in a particular district in some specified month, then dengue risk in its surrounding districts will increase with a peak in the next 12 months. Similar to Figure 8a this suggests a negative feedback cyclic pattern. For the sake of brevity, the lag-response curve for the differing number of dengue cases is not shown here. Thus, based on the aforementioned analyses, we use the optimal lag terms of dengue incidences in surrounding districts (lag 1, lag 2 and lag 12) along with the optimal meteorological variables [2] and optimal lagged dengue variables (used in model D). This new model has the representation of all the optimal variables and hence termed as 'Optimal Representation Model' (model E).

The association of these terms with dengue transmission is shown in Figure 10. It is observed that for short-term lagged data of dengue cases of surrounding districts the association with dengue transmission in the target district is almost linear. For lag 1, transmission

[2]The association with rainfall data for the current month (i.e. lag 0) was found to be statistically insignificant ($p \gg 0.05$) for this model.



(a)



(b)

Figure 8: Retrospective transmission period is calculated to account for the influence of dengue incidences in each of the target districts (Fig 8a) and lag-response curve for an increase in 1100 units of dengue incidences (Fig 8b).

increases while for lag 2 it decreases. However, for long-term lagged data (lag 12) the association slightly decreases for a small number of dengue cases (up to about 75 cases) and then modestly increases before being constant. The performance of the model using the training data from 2008 – 2012 calculated using R-squared (adjusted) is 0.71, RMSE is 5.55 and SRMSE is 0.34. The 'Deviance Explained' is 72.0%. The visual analyses of the prediction of the 'Optimal Representation' model is shown in Figure 11e.

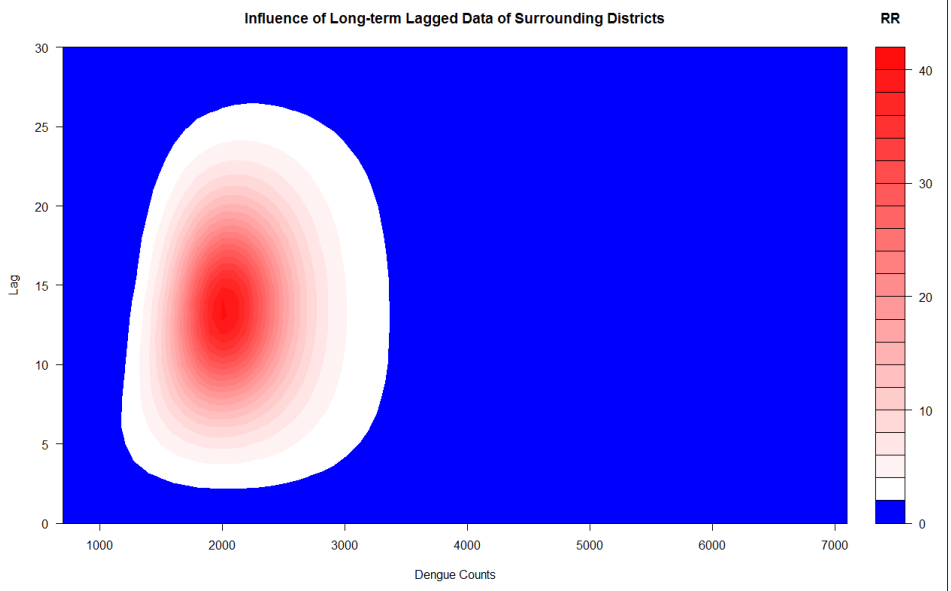


Figure 9: Contour plot of the relative risks of transmission of dengue in a target district up to 30 months following a larger outbreak of dengue in its surrounding districts in a specified month.

The vertical axis in Figure 11 represents the number of DHF cases and the horizontal axis depicts the time in months from 2008-2012. Figure 12 shows the influence of socioeconomic data on prediction of dengue counts. The performance of1the model using the training data from 2008 – 2012 calculated using R-squared (adjusted) is 0.77, RMSE is 5.54 and SRMSE is 0.34. The 'Deviance Explained' is 77.0%. The visual analyses of the prediction of the 'Socioeconomic data Included' model are shown in Figure 12b. The analyses show that the relative risks of dengue transmission almost linearly decreases as the garbage collection increases. It might be because the increase in garbage collection by municipal authorities is spatially correlated with higher economic activity in a region, which may signify better sanitation and healthier lifestyle[59].

4 Discussion

As discussed in our analyses above, the 'Optimal Representation Model' (model E) shows the best[3] performance of all the aforementioned models (model A-F). The predictive

[3]Although it is shown in Table 1 that model F has similar predictive performance as that of model E, since the later has comparatively smaller complexity it is considered to be the candidate model for evaluation on external data set. However, in terms of over-all performance, model F is deemed to be most suitable.

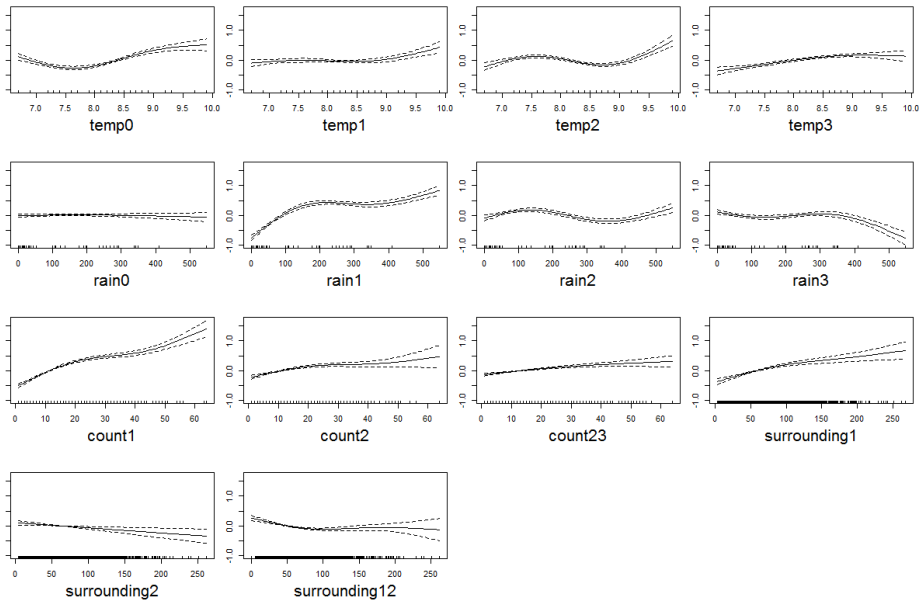


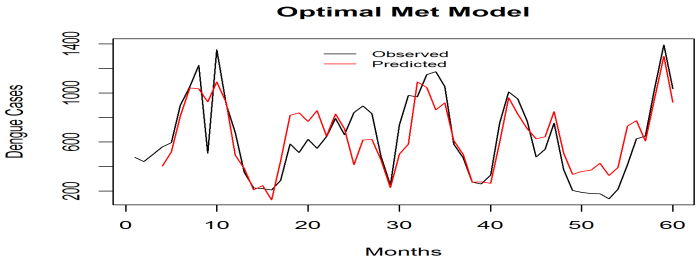
Figure 10: **The optimal model:** Association between the meteorological variables, past dengue count over optimal lags within 1-30 months, surrounding district count over 0–30 months and the dengue outbreak. Solid lines represent relative risks (RR) of dengue cases whereas the dotted lines depict the upper and lower limits of 95% confidence intervals.

Model Name	RMSE	SRMSE	R-sq.(adj)	Deviance Explained
A: Meteorology Optimal	8.46	0.52	0.28	0.32
B: Short-term Lag Surveillance Model	7.90	0.49	0.40	0.41
C: Optimal Lag Surveillance Model	7.07	0.43	0.53	0.53
D: Optimal Met and Lag Surveillance Model	5.72	0.35	0.69	0.70
E: Optimal Representation Model	5.55	0.34	0.71	0.72
F: Social-economic data Included	5.54	0.34	0.71	0.77

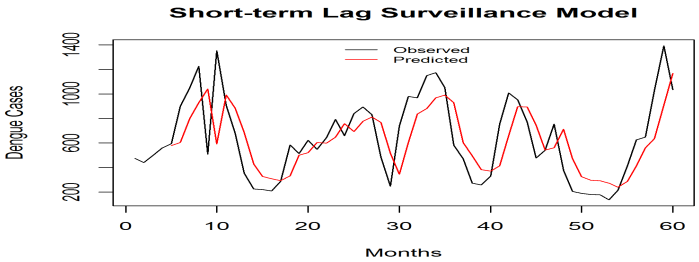
Table 1: Predictive performance statistics of different models evaluated on the training data. The performance is measured on different metrics. The best model should have the lowest errors (RMSE, SRMSE) and have the best fit (measured in R-sq.(adj).)

performance statistics of different models evaluated on the training data from the years $2008-2012$ is also listed in Table 1.

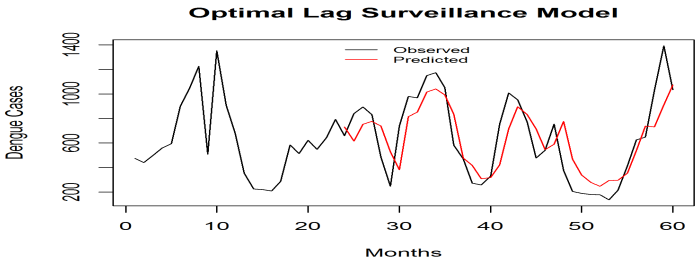
The simplest model (A) using only meteorological variables showed the poorest performance when compared with the other models. The quality of fit and predictive ability increases with the subsequent models as shown (Table 1). For instance, the model B that includes short-term dengue incidences data, showed, poor predictions as compared to the model C which has optimal autoregressive model terms including dengue at lag 1, 2 and lags 23. The combination of optimal meteorological variables with optimal autoregressive model terms had a rather good predictive performance. Finally, the combination of optimal autoregressive model terms of the data from surrounding districts has the best predictive ability. The predictive ability as evaluated by RMSE and SMRSE, as well as the values of R-sq.(adjusted) and 'Deviance Explained' for the models A-E show that model (E) combining model (A) and (D) is the best-predictive model (Table 1).



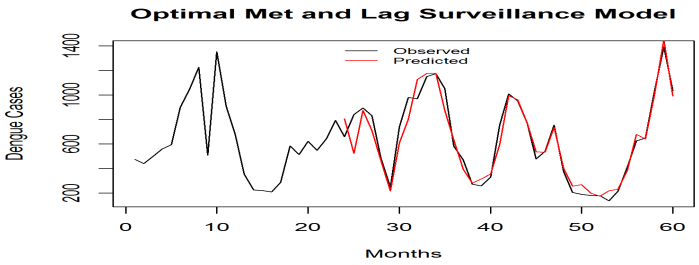
(a) Only Rainfall and DTR data with statistical significance (past 4 months)was used.



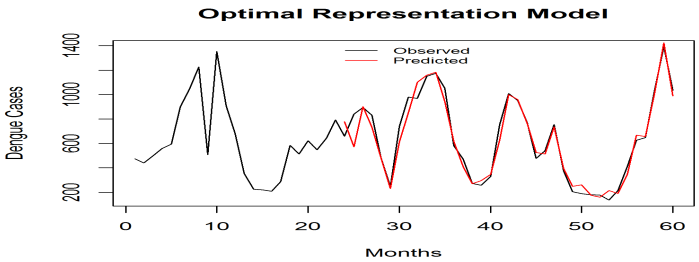
(b) The data of dengue counts of districts from optimal short-lag (i.e. past 2 months) was used.



(c) Along with the dengue data of optimal short-lag, the long-lag (24th month) was also used.

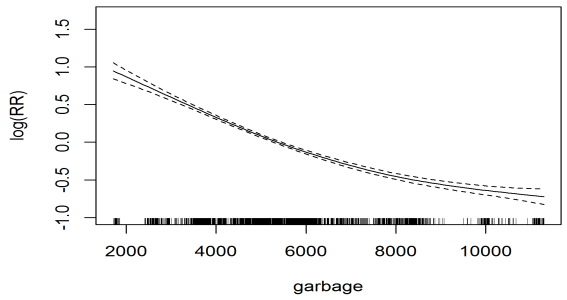


(d) The optimal data (rainfall, DTR, lagged dengue data) of each target district is used.

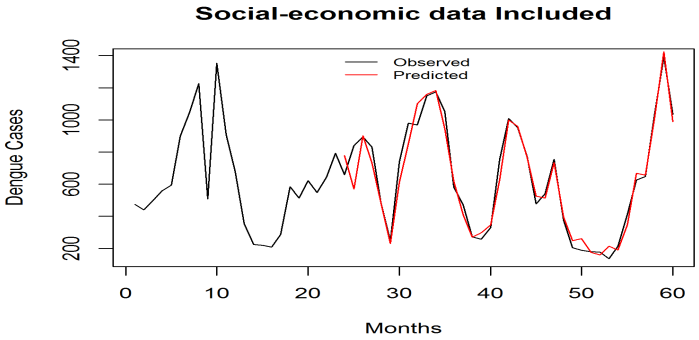


(e) The optimal combination of data (rainfall, DTR, lagged dengue data) of each target district and its *surrounding district* is used.

Figure 11: **Monthly Observed and Predicted Dengue Cases from 2008-2012.** Black line represent the observed DHF cases and red line represents the predicted cases.



(a) The the smooth term for representing the variable for monthly garbage collection.



(b) The optimal combination of data (rainfall, DTR, lagged dengue data) of each target district and its *surrounding district* as well the data representing social capital i.e. monthly garbage collection is used.

Figure 12: Influence of socioeconomic data on prediction of dengue counts.

The final "Optimal Representation" model (E) included combinations of the meteorological variables and the autoregressive lag terms of dengue counts in the past (both for target districts and its surrounding districts) according to:

$$log(C_{0,t}) \sim \alpha + \sum_{l=0}^{3} ns(\mathfrak{T}_{lt}, d=3) + \sum_{l=0}^{3} ns(R_{lt}, d=3) + \sum_{l=1}^{2} ns(C_{lt}, d=3) + \sum_{l=23}^{23} ns(C_{lt}, d=3) + \sum_{l=1}^{2} ns(SC_{lt}, d=3) + \sum_{l=12}^{12} ns(SC_{lt}, d=3) \quad (5)$$

where C represents the dengue counts in the target district, $\mathfrak{T}$ represents DTR (°C), R denotes the mean monthly rainfall (mm) and SC denotes the dengue counts in the surrounding districts.

After the 'Optimal Representation' model of Equation 5 is fit, the residuals are inferred; and their normality and residual autocorrelation are checked. Figure 13 shows that residual histograms (top-left) are symmetric and follow a unimodal distribution. The Q-Q plot of deviance residuals which are conditional on the fitted model coefficients and scale parameter (bottom-left) is close to a straight line. This suggests that the distributional assumptions

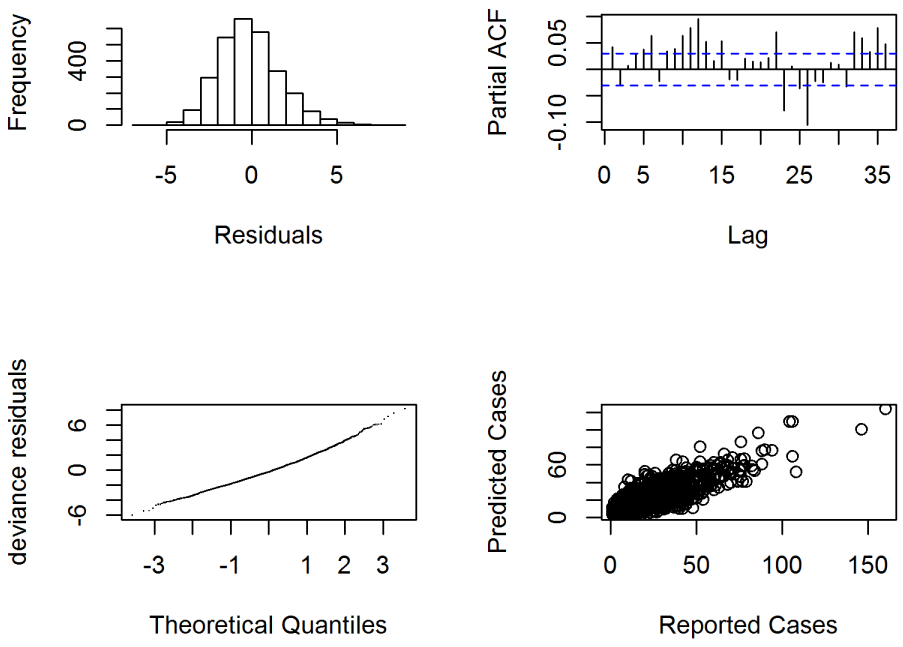


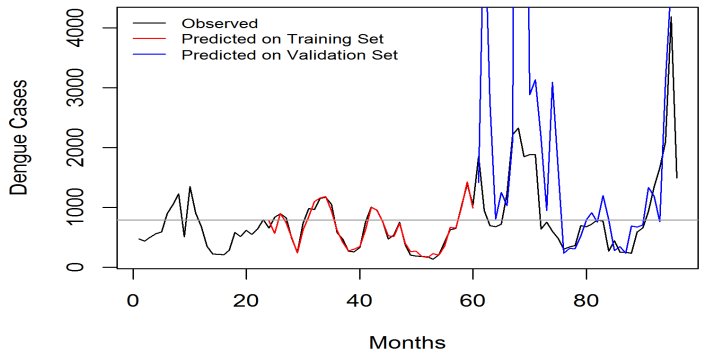
Figure 13: **Residual Diagnosis.** Top-left is the histogram of residuals; top-right is the partial ACF plot. On bottom-left is the Q-Q plot for the deviance residuals whereas the relationship between the reported and predicted cases is shown on bottom-right

.

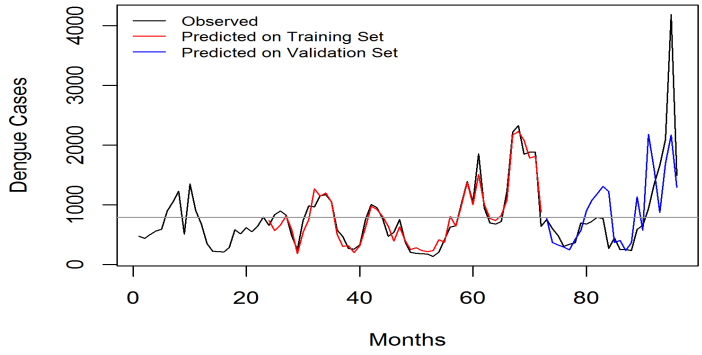
of the model are satisfied. The plot of reported and predicted cases indicate linear relation (bottom-right). The partial autocorrelation function plot (top-right) does show significant autocorrelation for long-term lags (lag $10-13$, lag 22, lag 25 etc.), but it was found that keeping those lagged terms increases the complexity of the model without any significant increase in prediction performance. Thus, they were chosen to be ignored in the model [4].

To evaluated the predictive performance of our final model (E), several external validation data sets were created. The visual analyses is provided in Figure 14 and the results are mentioned in Table 2. The 'red' line shows the prediction on training data set and the 'blue' line shows the prediction on validation sets while the 'black' line represents the observed dengue cases. The horizontal gray line represents the mean monthly dengue cases across $2008-2015$ which we consider to be the threshold of the epidemic. The visual analysis of Figure 14a shows the poorest performance with a significant difference between predicted and observed cases for the external data set (i.e. for the months in the year 2013 and so on). The model trained on the larger set of training data shows the better predictive performance as the influence of both, the direct and retrospective transmission could be learned in the model (Table 2).

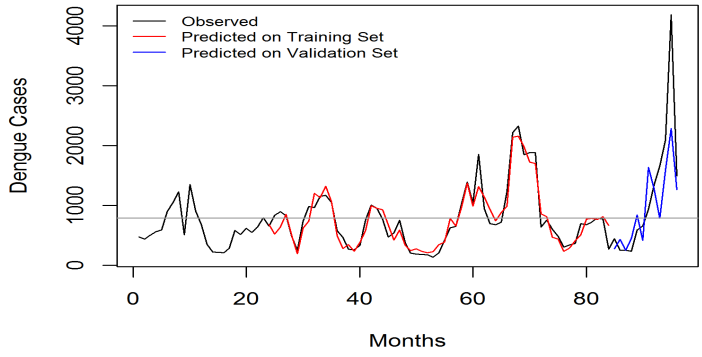
[4]Including those additional terms does not as well provide some additional insight in favor of our hypothesis too



(a) The model **E** (Equation 5) was trained on the data from 2008 – 2012 for validation on the external data set.



(b) The model **E** (Equation 5) was trained on the data from 2008 – 2013 for validation on external data set.



(c) The model E (Equation 5) was trained on the data from 2008 – 2014 for validation on external data set.

Figure 14: **Predicted Dengue Cases Versus Reported Dengue Cases in 2008-2015.** Black line represent the observed DHF cases and red line represents the predicted cases. The blue line represents the predicted cases on the external data set for which the model is not fit. The grey horizontal line represents the epidemic threshold which is the mean monthly DHF count across 2008-2015. The vertical axes represents the number of DHF cases and the horizontal axes represents the time in months from 2008-2015.

The discriminating ability of model E was evaluated against our specified threshold of the epidemic (790 monthly dengue cases) and WHO threshold of the epidemic. To separate transmission months above and below 790 cases, the specificity was 85.1% and sensitivity was 84.7%. The positive predictive value was observed to be 91.6% and the negative predictive value was 74.5%. The WHO moving outbreak threshold [60] is the sum total of monthly average cases of preceding 5 years and twice their standard deviation. This value comes out to be around 1287 dengue cases per month. To separate transmission months above and below 1287 cases, the specificity was 72.72% and sensitivity was 96.0%. The positive predictive value and negative predictive value were both estimated to be around 89.0%.

Training Dataset	**In-sample**	**Out-Sample** (2013-2015)	**Out-Sample** (2014-2015)	**Out-Sample** (2013)	**Out-Sample** (2014)	**Out-Sample** (2015)
2008-2012	0.37	303.28	410.74	21.66	1.43	437.50
2008-2013	0.38		0.56		0.32	0.54
2008-2014	0.39					0.49

Table 2: Predictive performance statistics of most optimal model evaluated on validation data sets is measured in SRMSE.

5 Conclusion

In this study, the dengue incidences were predicted using a variety of data. The best model for dengue prediction includes meteorological data (rainfall, DTR), lagged dengue data from specified districts and their surroundings and the socioeconomic data. We proved that for the prediction of dengue outbreaks within a district, the influence of dengue incidences and socioeconomic data from the surrounding districts is statistically significant. Thus for forecasting dengue outbreaks and taking preventing measures, the epidemiologists and health authorities should consider the influence of movement patterns of people and spatial heterogeneity of human activities. In our future work, we intend to develop customized models for each individual district, include more demographic data and data from government surveys e.g. House index (HI), Breteau index (BI) and Container index (CI). Also, we would like to integrate the information from social media platforms to track the dengue incidences in real time.

List of abbreviations

DENV: Dengue Virus; DF: Dengue Fever; DHF: Dengue Haemorrhagic Fever; DSS: Dengue Shock Syndrome; DTR: Diurnal Temperature Range; GAM: Generalized Additive Model; DLNM: Distributed Lag Non-Linear Models; RMSE: Root Mean Squared Error; SRMSE: Standard Root Mean Squared Error.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
No funds were obtained to carry out this project.

Author's contributions
RJ analyzed and interpreted the data. He also wrote the manuscript. SS collected and cleaned the data. SI helped make data acquisition task easier by coordinating between different sources. HP supervised the project. All authors read and approved the final manuscript.

Acknowledgements
Mrs. Sarinthorn Sontisirikit, Senior Public Health Officer of the Department of Disease Control) enabled surveying in the local communities of Kannayao, Lat Krabang , Thung khru, Saphansung and Jatujak districts of Bangkok.

Author details

[1]National Institute of Informatics, Tokyo, Japan. [2]Asian Institute of Technology, School of Engineering and Technology, Bangkok, Thailand. [3]Department of Disease Control Thirteenth Division, Bangkok, Thailand.

References

1. Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., *et al.*: The global distribution and burden of dengue. Nature **496**(7446), 504–507 (2013)
2. Guzmán, M.G., Kouri, G.: Dengue: an update. The Lancet infectious diseases **2**(1), 33–42 (2002)
3. Gubler, D.J., Clark, G.G.: Dengue/dengue hemorrhagic fever: the emergence of a global health problem. Emerging infectious diseases **1**(2), 55 (1995)
4. WHO: Dengue Type. http://www.who.int/mediacentre/factsheets/fs117/en/. Accessed 2015-11-20 (2015)
5. Hammon, W.M., Sather, G.: Virological findings in the 1960 hemorrhagic fever epidemic (dengue) in thailand. The American journal of tropical medicine and hygiene **13**(4), 629–641 (1964)
6. Chareonsook, O., Foy, H., Teeraratkul, A., Silarug, N.: Changing epidemiology of dengue hemorrhagic fever in thailand. Epidemiology and infection **122**(01), 161–166 (1999)
7. Hesse, R.R.: Dengue virus evolution and virulence models. Clinical Infectious Diseases **44**(11), 1462–1466 (2007)
8. Wilder-Smith, A., Renhorn, K.-E., Tissera, H., Bakar, S.A., Alphey, L., Kittayapong, P., Lindsay, S., Logan, J., Hatz, C., Reiter, P., et al.: Denguetools: innovative tools and strategies for the surveillance and control of dengue. Global health action **5** (2012)
9. Leitmeyer, K.C., Vaughn, D.W., Watts, D.M., Salas, R., Villalobos, I., Ramos, C., Rico-Hesse, R., *et al.*: Dengue virus structural differences that correlate with pathogenesis. Journal of virology **73**(6), 4738–4747 (1999)
10. Runge-Ranzinger, S., Horstick, O., Marx, M., Kroeger, A.: What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? Tropical Medicine & International Health **13**(8), 1022–1041 (2008)
11. Thammapalo, S., Chongsuvivatwong, V., Geater, A., Dueravee, M.: Environmental factors and incidence of dengue fever and dengue haemorrhagic fever in an urban area, southern thailand. Epidemiology and Infection **136**(01), 135–143 (2008)
12. Sarfraz, M.S., Tripathi, N.K., Kitamoto, A.: Near real-time characterisation of urban environments: a holistic approach for monitoring dengue fever risk areas. International Journal of Digital Earth **7**(11), 916–934 (2014)
13. Scott, T.W., Morrison, A.C.: Aedes aegypti density and the risk of dengue-virus. Ecol. Aspects Appl. Genet. Modi. Mosq **2**, 187 (2003)
14. Alto, B.W., Lounibos, L.P., Mores, C.N., Reiskind, M.H.: Larval competition alters susceptibility of adult aedes mosquitoes to dengue infection. Proceedings of the Royal Society of London B: Biological Sciences **275**(1633), 463–471 (2008)
15. Syarifah, N., Rusmatini, T., Djatie, T., Huda, F.: Ovitrap ratio of aedes aegypti larvae collected inside and outside houses in a community survey to prevent dengue outbreak, bandung, indonesia, 2007. Proc Assoc Southeast Asian Nations Congr Trop Med Parasitolol **3**, 116–120 (2008)
16. Chareonviriyaphap, T., Akratanakul, P., Nettanomsak, S., Huntamai, S.: Larval habitats and distribution patterns of aedes aegypti (linnaeus) and aedes albopictus (skuse), in thailand. (2003)
17. Raju, K., Sokhi, B.: Application of gis modeling for dengue fever prone area based on socio-cultural and environmental factors–a case study of delhi city zone. Int Arch Photogramm Remote Sens Spat Inf Sci **37**, 165–170 (2008)
18. Favier, C., Schmit, D., Müller-Graf, C.D., Cazelles, B., Degallier, N., Mondet, B., Dubois, M.A.: Influence of spatial heterogeneity on an emerging infectious disease: the case of dengue epidemics. Proceedings of the Royal Society of London B: Biological Sciences **272**(1568), 1171–1177 (2005)
19. Arunachalam, N., Tana, S., Espino, F., Kittayapong, P., Abeyewickrem, W., Wai, K.T., Tyagi, B.K., Kroeger, A., Sommerfeld, J., Petzold, M.: Eco-bio-social determinants of dengue vector breeding: a multicountry study in urban and periurban asia. Bulletin of the World Health Organization **88**(3), 173–184 (2010)
20. Chang, A.Y., Parrales, M.E., Jimenez, J., Sobieszczyk, M.E., Hammer, S.M., Copenhaver, D.J., Kulkarni, R.P.: Combining google earth and gis mapping technologies in a dengue surveillance system for developing countries. International journal of health geographics **8**(1), 49 (2009)
21. Knudsen, A.B., Slooff, R.: Vector-borne disease problems in rapid urbanization: new approaches to vector control. Bulletin of the World Health Organization **70**(1), 1 (1992)
22. Troyo, A., Fuller, D.O., Calderón-Arguedas, O., Solano, M.E., Beier, J.C.: Urban structure and dengue incidence in puntarenas, costa rica. Singapore journal of tropical geography **30**(2), 265–282 (2009)
23. Nakhapakorn, K., Tripathi, N.K.: An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. International Journal of Health Geographics **4**(1), 13 (2005)
24. World Population Review: Bangkok Population 2015. http://worldpopulationreview.com/world-cities/bangkok-population/. Accessed 2016-01-18 (2015)
25. Stoddard, S.T., Wearing, H.J., Reiner Jr, R.C., Morrison, A.C., Astete, H., Vilcarromero, S., Alvarez, C., Ramal-Asayag, C., Sihuincha, M., Rocha, C., *et al.*: Long-term and seasonal dynamics of dengue in iquitos, peru. PLoS Negl Trop Dis **8**(7), 3003 (2014)
26. Lambrechts, L., Paaijmans, K.P., Fansiri, T., Carrington, L.B., Kramer, L.D., Thomas, M.B., Scott, T.W.: Impact of daily temperature fluctuations on dengue virus transmission by aedes aegypti. Proceedings of the National Academy of Sciences **108**(18), 7460–7465 (2011)
27. Carrington, L.B., Seifert, S.N., Armijos, M.V., Lambrechts, L., Scott, T.W.: Reduction of aedes aegypti vector competence for dengue virus under large temperature fluctuations. The American journal of tropical medicine and hygiene **88**(4), 689–697 (2013)
28. Sulaiman, S., Pawanchee, Z.A., Arifin, Z., Wahab, A.: Relationship between breteau and house indices and cases of dengue/dengue hemorrhagic fever in kuala lumpur, malaysia. American Mosquito Control Association **12**, 494–496 (1996)
29. Halstead, S.B.: Dengue virus-mosquito interactions. Annu. Rev. Entomol. **53**, 273–291 (2008)
30. Esteva, L., Vargas, C.: Influence of vertical and mechanical transmission on the dynamics of dengue disease. Mathematical biosciences **167**(1), 51–64 (2000)
31. Adams, B., Holmes, E., Zhang, C., Mammen, M., Nimmannitya, S., Kalayanarooj, S., Boots, M.: Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in bangkok. Proceedings of the National Academy of Sciences **103**(38), 14234–14239 (2006)

32. Cummings, D.A., Iamsirithaworn, S., Lessler, J.T., McDermott, A., Prasanthong, R., Nisalak, A., Jarman, R.G., Burke, D.S., Gibbons, R.V., *et al.*: The impact of the demographic transition on dengue in thailand: insights from a statistical analysis and mathematical modeling. PLoS medicine **6**(9), 999 (2009)
33. WHO: Human Dengue Symptoms. http://www.who.int/denguecontrol/human/en/. Accessed 2015-11-19 (2015)
34. Gubler, D.J.: Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. Trends in microbiology **10**(2), 100–103 (2002)
35. Gubler, D.J.: Cities spawn epidemic dengue viruses. Nature medicine **10**(2), 129–130 (2004)
36. Anuradha, S., Singh, N., Rizvi, S., Agarwal, S., Gur, R., Mathur, M.: The 1996 outbreak of dengue hemorrhagic fever in delhi, india. (1998)
37. Vaughn, D.W.: Invited commentary: Dengue lessons from cuba. American Journal of Epidemiology **152**(9), 800–803 (2000)
38. Ooi, E.-E., Gubler, D.J.: Dengue in southeast asia: epidemiological characteristics and strategic challenges in disease prevention. Cadernos de saude publica **25**, 115–124 (2009)
39. Guzmán, M.G., Kourı' G.: Dengue diagnosis, advances and challenges. International journal of infectious diseases **8**(2), 69–80 (2004)
40. Erlanger, T., Keiser, J., Utzinger, J.: Effect of dengue vector control interventions on entomological parameters in developing countries: a systematic review and meta-analysis. Medical and veterinary entomology **22**(3), 203–221 (2008)
41. Horstick, O., Runge-Ranzinger, S., Nathan, M.B., Kroeger, A.: Dengue vector-control services: how do they work? a systematic literature review and country case studies. Transactions of the Royal Society of Tropical Medicine and Hygiene **104**(6), 379–386 (2010)
42. Vanlerberghe, V., Toledo, M., Rodriguez, M., Gomez, D., Baly, A., Benitez, J., Van Der Stuyft, P.: Community involvement in dengue vector control: cluster randomised trial. Bmj **338**, 1959 (2009)
43. Luz, P.M., Vanni, T., Medlock, J., Paltiel, A.D., Galvani, A.P.: Dengue vector control strategies in an urban setting: an economic modelling assessment. The Lancet **377**(9778), 1673–1680 (2011)
44. Tatem, A.J., Hay, S.I., Rogers, D.J.: Global traffic and disease vector dispersal. Proceedings of the National Academy of Sciences **103**(16), 6242–6247 (2006)
45. World Weather Online: Bangkok average temperature and rain from year 2000 to 2012. http://worldpopulationreview.com/world-cities/bangkok-population/. Accessed 2016-01-18 (2012)
46. Johansson, M.A., Cummings, D.A., Glass, G.E.: Multiyear climate variability and dengue—el nino southern oscillation, weather, and dengue incidence in puerto rico, mexico, and thailand: a longitudinal data analysis. PLoS Med **6**(11), 1000168 (2009)
47. Morin, C.W., Comrie, A.C., Ernst, K.: Climate and dengue transmission: evidence and implications. Environmental Health Perspectives (Online) **121**(11-12), 1264 (2013)
48. Thai, K.T., Anders, K.L.: The role of climate variability and change in the transmission dynamics and geographic distribution of dengue. Experimental Biology and Medicine **236**(8), 944–954 (2011)
49. of Vector-Borne Diseases Thailand, B.: Dengue fever situation in Thailand. http://www.thaivbd.org. Accessed 2015-12-1 (2015)
50. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016). R Foundation for Statistical Computing. https://www.R-project.org/
51. Wood, S.N.: Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, ??? (2006)
52. Ramadona, A.L., Lazuardi, L., Hii, Y.L., Holmner, Å., Kusnanto, H., Rocklöv, J.: Prediction of dengue outbreaks based on disease surveillance and meteorological data. PloS one **11**(3), 0152688 (2016)
53. Hii, Y.L., Rocklöv, J., Wall, S., Ng, L.C., Tang, C.S., Ng, N.: Optimal lead time for dengue forecast. PLoS Negl Trop Dis **6**(10), 1848 (2012)
54. Reich, N.G., Shrestha, S., King, A.A., Rohani, P., Lessler, J., Kalayanarooj, S., Yoon, I.-K., Gibbons, R.V., Burke, D.S., Cummings, D.A.: Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. Journal of The Royal Society Interface **10**(86), 20130414 (2013)
55. Gasparrini, A., Armstrong, B., Kenward, M.G.: Distributed lag non-linear models. Statistics in medicine **29**(21), 2224–2234 (2010)
56. Mondini, A., Chiaravalloti-Neto, F.: Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a brazilian city. Science of the Total Environment **393**(2), 241–248 (2008)
57. Wu, P.-C., Lay, J.-G., Guo, H.-R., Lin, C.-Y., Lung, S.-C., Su, H.-J.: Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical taiwan. Science of the total Environment **407**(7), 2224–2233 (2009)
58. Gasparrini, A.: Distributed lag linear and non-linear models in r: the package dlnm. Journal of statistical software **43**(8), 1 (2011)
59. Satterthwaite, D.: Environmental transformations in cities as they get larger, wealthier and better managed. Geographical Journal, 216–224 (1997)
60. Organization, W.H., for Research, S.P., in Tropical Diseases, T., of Control of Neglected Tropical Diseases, W.H.O.D., Epidemic, W.H.O., Alert, P.: Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control. World Health Organization, ??? (2009)

Appendix

Table 3: Approximate significance of smooth terms depicting lagged meteorological data.

Results				
P-Value	**Lag 0**	**Lag 1**	**Lag 2**	**Lag 3**
Mean Monthly DTR	5.26e-16	0.000261	0.000921	2e-16
Cumulative Monthly Rainfall	8.96e-15	2e-16	2e-16	8.77e-10
R-sq(adj.)	0.283			
Deviance Explained	31.9%			
RMSE	8.462			
SRMSE	0.52			

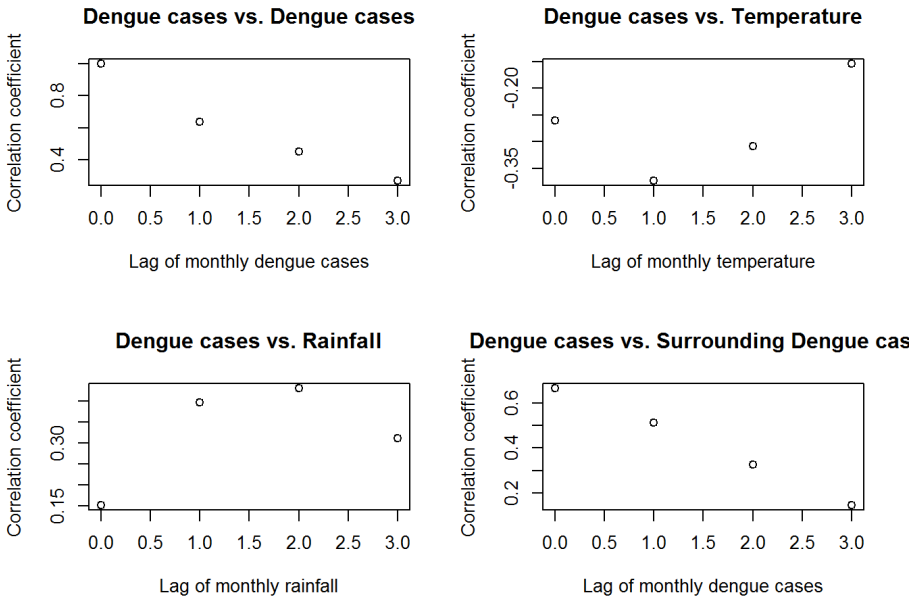


Figure 15: Cross-correlation between the Outcome and the Predictors.

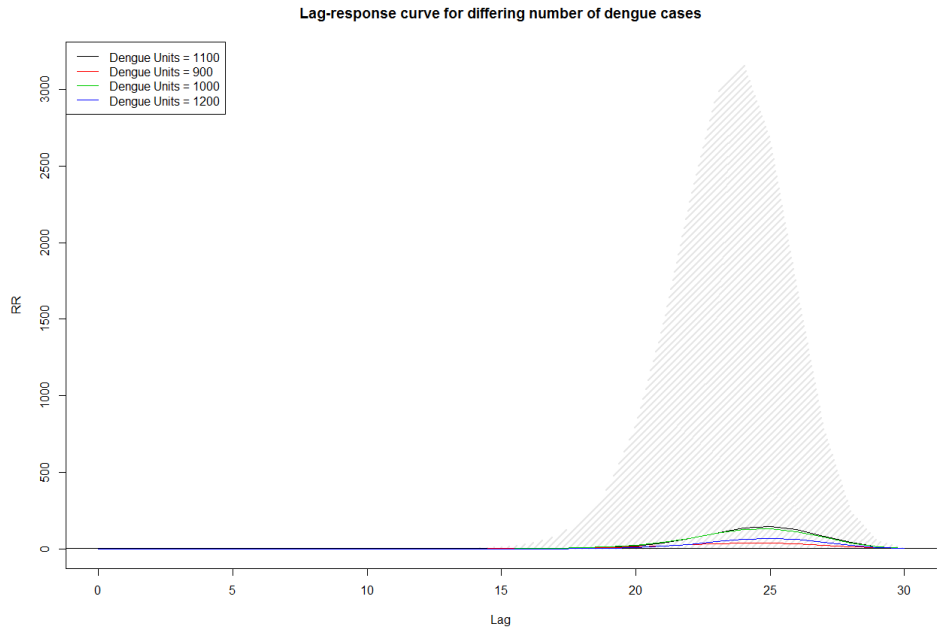


Figure 16: Association of different units of increase in the number of dengue cases with the specified number of months.

Table 4: Approximate significance of smooth terms depicting lagged meteorological data.

Results				
P-Value	**Lag 0**	**Lag 1**	**Lag 2**	**Lag 3**
Mean Monthly DTR	5.26e-16	0.000261	0.000921	2e-16
Cumulative Monthly Rainfall	8.96e-15	2e-16	2e-16	8.77e-10
R-sq(adj.)	0.283			
Deviance Explained	31.9%			
RMSE	8.462			
SRMSE	0.52			