Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India

Nikita Agarwal, Shiva Reddy Koti*, Sameer Saran and A. Senthil Kumar

Indian Institute of Remote Sensing, ISRO, Dehradun 248 001, India

Dengue is a hazardous disease which poses a critical threat to the population of Delhi, India. These cases are steadily reported during and post-monsoon season indicating its correlation with weather parameters. Establishing this relation will help understand the spread of dengue and will allow decision makers take precautionary steps beforehand. Our study explains the adopted multi-regression and Naïve Bayes approach to model the relation between dengue cases and weather parameters, i.e. maximum temperature, rainfall and relative humidity. Both these models have served a great deal in modelling this relationship which has enabled us to forecast a probable dengue outbreak. Our results have shown that sudden and high rainfall accompanied with 30–35°C temperature and high relative humidity contributes to a highly vulnerable weather for the spread of dengue. Also, we have proposed a new application of spherical *k*-means clustering algorithm to identify zones with similar transmission pattern which gives insight into the distribution of dengue incidences in Delhi. Results show that Central, Civil Lines, Rohini, South and West zones have the highest odds of dengue occurrences.

Keywords: Dengue, multi-regression, Naïve Bayes, spherical *k*-means, weather parameters.

DENGUE is an arboviral infection transmitted through mosquitoes namely, *Aedes aegypti* and *Aedes albopictus*. The primary vector of dengue, *Aedes aegypti*, is ordinarily found between 35°N and 35°S. The whole tropical area including America, Asia and Africa is prone to dengue, and highly populated countries like India, Indonesia, Brazil, China have the greatest burden of share[1]. Their survival rate is very low during the colder months of winter. Hence, they are found near 45°N only during warmer months[2]. These mosquitoes originally domesticated from Africa, and can easily colonize by laying eggs wherever they can find stagnant water[3]. According to Gubler[4], during and after the Second World War, these mosquitoes became pervasive in Southeast Asia bringing a dengue pandemic to the world.

*For correspondence. (e-mail: shivareddy@iirs.gov.in)

Dengue in India was first reported in or around 1946 and today, impacts 20,474 lives on an average[5,6]. Northern and Eastern cities of India and the whole of the Gangetic Plain are plagued with *Aedes aegypti*, the primary vector of dengue[7]. In southern India, however, the secondary vector, *Aedes albopictus* is more prevalent[8]. Many Indian states including Andhra Pradesh, Delhi, Gujarat, Goa, Haryana, Karnataka, Kerala, Maharashtra, Rajasthan, Uttar Pradesh, Puducherry, Punjab, Tamil Nadu, West Bengal and Chandigarh have seen dengue outbreak and in fact, according to the World Health Oranization, cyclic epidemics are increasing in India[2,5,9].

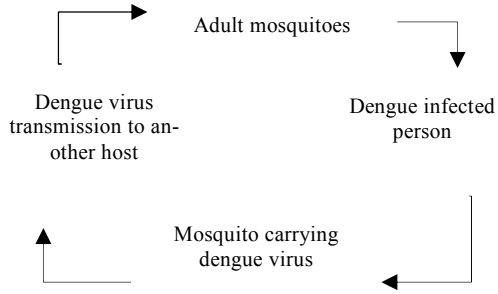
Delhi, in particular, is more vulnerable than other cities due to a large number of asymptomatic dengue case occurrences[10]. People of Delhi lack immunity to the dengue virus despite the historical outbreaks during 1967, 1970, 1982, 1988 and 1996 and in the recent past during 2006, 2010, and 2013 (refs 5, 10, 11). These outbreaks are normally reported during or post-monsoon season signifying their relation with weather parameters.

An attempt was made to identify the relation between dengue and weather parameters, i.e. temperature, rainfall and relative humidity and use this derived relation to forecast dengue vulnerability. Multi-variate regression and Naïve Bayes approach was adopted to study the relation between weather and dengue. Also, we report a new application of spherical *k*-means clustering algorithm to group the zones of Delhi with similar transmission patterns.

Related studies

An understanding is developed about the characteristics of *Aedes aegypti*, i.e. its behaviour, habitat preference, transmission pattern, etc. from various clinical literature[2,8,12,13] (Table 1). Apart from the facts in Table 1, it is worth noting that temperature fluctuation, like high diurnal temperature, can reduce the extrinsic incubation period (EIP) of mosquitoes and accelerate the dengue virus transmission[13]. At low diurnal temperature, mosquito infection and its transmission are raised and at high diurnal temperature, infections are reduced[14]. These daily and seasonal changes also define the geographical limits

Table 1. Information on *Aedes aegypti* and dengue

Category name	Description	Category value
Oviposition preference	Preference for laying eggs	Containers commonly found in household.
Larval habitat		Tree holes, coconut shells, leaf axils
Extrinsic incubation period	The time it takes for a mosquito to transmit virus to host	8–10 days
Intrinsic incubation period	The time it takes for a human to transmit virus to a mosquito	3–10 days
Dengue virus transmission pattern	Transmission cycle that takes place from mosquito to human and human to mosquito.	
Symptoms of dengue		Sudden fever, headaches, severe joint and muscle pains, fatigue, vomiting, skin rash, pain behind eyes, nausea, mild bleeding, etc.
Miscellaneous		These mosquitoes are rare above 1000 m from mean sea level. Eggs can survive without water for one year

of *Aedes aegypti* and *Aedes albopictus* thus limiting the geographical transmission of the dengue virus[15].

There have been a few empirical studies to determine the correlation between dengue and weather parameters, most of which were conducted in Singapore, Malaysia and Thailand. These studies have used different techniques, e.g. multi-variate regression, Poisson regression, logistic regression and classification.

One such study applied multi-regression to find the correlation between dengue cases and weather parameters[16]. The empirical model was created by taking weather parameters, viz. temperature, rainfall and humidity as the independent variable and dengue cases as the dependent variable. The developed model serves to denote the correlation between weather parameters and dengue cases. Apart from these weather parameters, the authors have tried to establish a relation with land use/land cover. Using Bayes' theorem, they extracted *i*-value (information value) between dengue cases and land use/land cover type. According to these calculated *i*-values, the built-up area is found to be a highly risky zone and the forest area showed the lowest risk.

A similar study in Singapore adopted Poisson regression model[17]. The authors have modelled the relative risk of various lag periods. They reported that higher weekly mean temperature and increased precipitation led to higher number of dengue cases during the years 2004–2007. The study incorporated factors which cannot be explained by weather parameters, i.e. vector control capacity, heard immunity, change of dengue serotype by a trend function.

Depending on the regions and their diverse weather conditions, the relation of weather parameters with dengue deviates[18]. In a study carried out near the bordering area of the Andaman Sea and the Gulf of Thailand, a preliminary test was performed to identify the correlation between dengue incidences and weather parameters, i.e. minimum temperature, maximum temperature, relative humidity and mean temperature. The study compared both regions in terms of weather parameters and tried to establish a relationship between deviations in weather conditions with irregularities in dengue incidences. That study indicated that rainy days and precipitation during monsoon season may as well play a significant role in spreading disease. These two parameters allow a sufficient window for mosquito eggs to hatch and turn into larvae and subsequently adult mosquitoes. Also, the study found that near the seaside, relative humidity seemed to have a positive effect on mosquito breeding sites, whereas it had a negative effect near the Gulf region.

A regression study in Dhaka (Bangladesh) included the average maximum temperature, average monthly rainfall and humidity[19]. The derived regression model was retrospectively validated. However, the study mentions that other parameters like immunological, entomological and demographical parameters can also be included along with weather parameters.

Data mining techniques are becoming popular to detect dengue outbreak[20]. A few studies have developed a predictive model for dengue outbreak[21]. The authors have used non-clinical data namely, year, epidemic week, age, gender, address, type of dengue, incubation period, death

code, repetition case, type of outbreak, etc. The studies have used multiple rule base classifier to detect dengue outbreak. They have used *k*-means clustering technique to select the rules created from various classification techniques. Weka data mining tool was also used to classify dengue cases[22,23]. Classification techniques, i.e. Naïve Bayes, J48 tree, SMO, Random Tree, etc. are used for prediction using symptoms as parameter.

Materials and methods

Study area

The study area chosen for the study is the National Capital Territory (NCT) of Delhi (India) as shown in Figure 1. It lies within 28.89°N and 76.82°E to 28.39°N and 77.34°E with approximately 1483 km^2. The elevation range is 200–253 m. The population of Delhi is approximately 16,314,838. The average yearly rainfall is 714 mm. The temperature can go up to 40–45°C and cools down to 4–5°C.

Datasets

We collected data on the monthly dengue cases from the Government of NCT. For the geospatial study we used the zone-wise distribution of dengue cases, collected from National Vector Borne Disease Control Programme (NVBDCP), Government of NCT[24]. Monthly weather parameters, i.e. rainfall, relative humidity and temperature were collected from the Statistical Abstract of Delhi 2014, Government of NCT of Delhi and the India Meteorological Department (IMD). The temperature, relative humidity and rainfall data of 10 automated weather stations (AWS) of Delhi were collected from IMD. The ward-wise population data was collected from the census

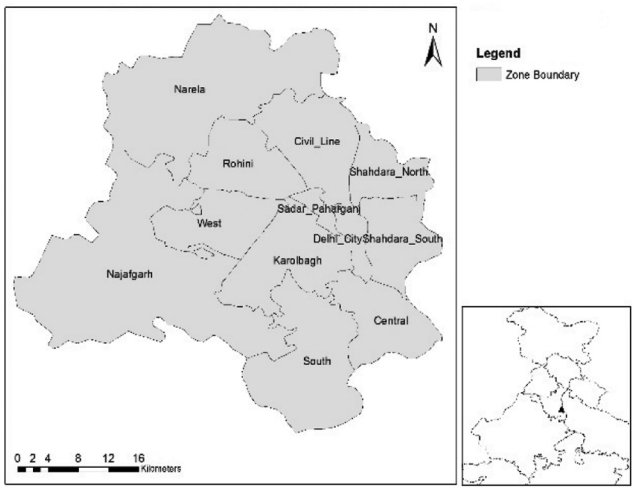


Figure 1. Study area.

information. The ward boundaries were matched with the zone boundaries in NVBDCP maps and zone-wise population was calculated. Table 2 gives the details of the datasets used in the study.

Methodology

Our entire methodology is divided into three approaches: multi-variate regression analysis, Naïve Bayes approach and spherical *k*-means clustering. As the dataset consists of numerical values, the regression gives the number of cases that can occur given the weather conditions. Based on historical data, the data is discretized into outbreak and non-outbreak for Naïve Bayes analysis.

Multi-variate regression analysis

Multi-variate regression analysis is an approach to model a relationship between dependent and independent variables. It overlaps statistical and machine learning aspects. Following is the characteristic equation of multivariate regression analysis with y as dependent variable, $x_1, x_2, \ldots, x_n$ as independent variables, $k_1, k_2, \ldots, k_n$ as coefficient of x and c as intercept.

$$y = k_1 * x_1 + k_2 * x_2 + \cdots + k_n * x_n + c.$$

Using this method we intend to build an empirical model from the data and use this model for prediction and forecast. In our scenario dengue incidences are dependent variables and weather parameters, i.e. maximum temperature and rainfall are independent variables. Due to high correlation and co-linearity among maximum temperature, minimum temperature and average temperature, the latter two are excluded from the independent variable set[19]. As our objective is to identify an outbreak, we are interested in peaks and dips over seasons. Hence before supplying data as input, spline smoothing was applied on data to smoothen irregularities within a season. Since it may take around 45 days for a mosquito to complete its life cycle and reach the adult stage, a lag of 2 months was projected between dengue cases and weather parameters[16]. First, we calculated maximum temperature, rainfall and fraction of cases in each zone. We collected the monthly averages of temperature, rainfall in the entire study area from the Statistical Abstract of Delhi, 2014 (SAD). In each zone, inverse distance weightage (IDW) and Thiessen Polygon were used for calculating maximum temperature and rainfall respectively. Since our temporal resolution for dengue incidence was monthly, we computed daily maximum temperature which was averaged over the entire month to get the monthly maximum temperature. These calculations were verified from the reported SAD values. The AWS data was used to calculate these weather parameters per zone. As SAD

Table 2. Description of dataset used in study

Variable	Unit	Time	Description	Source
Monthly dengue incidences	No. of people	January 2006–September 2015	Number of cases reported in Delhi at given time.	Govt of NCT of Delhi
Yearly zone wise dengue incidences (9 zones out of 12)	No. of people	2009–2013, 2015	Number of cases is various zones of Delhi	NVBDCP
Hourly temperature	°C	August 2011–October 2011; August 2012–October 2012; August 2014–October 2014; August 2015–October 2015	Hourly temperature of 10 AWS in Delhi	Indian Metrological Department
Hourly rainfall	mm		Rainfall amount at every hour of 10 AWS in Delhi	
Hourly relative humidity	%		Hourly relative humidity of 10 AWS in Delhi	
Maximum temperature	°C	January 2006–September 2015	Monthly average of daily maximum temperature	Statistical abstract of Delhi 2014, Government of NCT of Delhi
Minimum temperature	°C		Monthly average of daily minimum temperature	
Relative humidity	%		Average monthly relative humidity	
Total rainfall	mm		Total monthly rainfall	

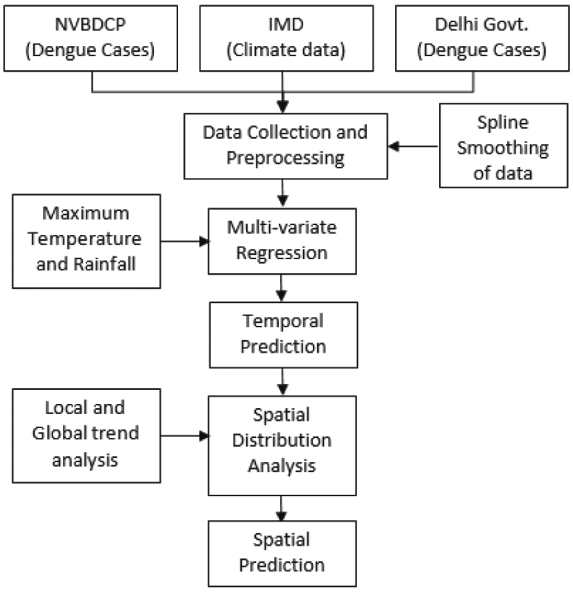


Figure 2. Methodology.

contained values for the entire National Capital Region (NCR), we averaged the values in each zone and compared those values to verify our computations. An empirical model was developed for temporal prediction from the monthly data of NCR obtained from SAD. This model was further refined to predict the spatial distribution of dengue incidences in Delhi zones. The refined model has two components: global trend and local trend. The global trend was calculated from the total number of cases predicted for Delhi by multiplying this number and fraction of cases for each zone. A fraction of cases was calculated as per the ratio of cases in each zone to total cases in Delhi. The local trend was calculated from zone-wise weather data using the developed empirical model. The empirical model was multiplied by the fraction of cases to get the local trend. Finally, the average of global and local components was taken. Figure 2 gives the schematic diagram of the aforementioned method.

Naïve Bayes Approach

The Naïve Bayes technique is a data mining technique which belongs to the probabilistic classifier. Naïve Bayes is widely used in text classification tasks and it assumes that the attributes are distributed independently[25]. It is based on Bayes' theorem which is

$$p(C_k \mid x) = \frac{[p(c_k)p(x \mid C_k)]}{[p(x)]} \quad (1)$$

where $x = (x_1, \ldots, x_n)$, $C = k$ possible classes, $p(x)$ = probability of x, $p(C_k)$ = probability of class k, $p(C_k|x)$ = conditional probability of class k for given x, and $p(x|C_k)$ = conditional probability of x for given class k.

The dataset is divided into training and test datasets to detect outbreak. Maximum temperature, rainfall, relative humidity and population density are used as attributes. The calculation of zonewise weather data is as explained in the previous section. From the historical dengue cases in NCR, the outbreak is defined as 97 or more cases per a million population. According to this definition, training dataset has been discretized and given as input to Naïve Bayes. The Naïve Bayes technique is chosen due to its efficiency of parameter estimation from small datasets[26].

Spherical k-means clustering

Spherical k-means, commonly used for clustering text documents, uses cosine distance for proximity measure to

cluster directional data[27]. A case vector can be formulated for each zone from the temporal data of dengue case incidences on which spherical *k*-means clustering can be applied. The idea is that the case vector for each zone gives a particular direction. Cosine similarity measure, when applied to these vectors, can give the distance between these vectors. Using this, distance matrix similarity between zones can be calculated. Using this similarity, these zones can be grouped according to their transmission patterns. The purpose of this clustering study is to identify if the zones have any similarity amongst each other in terms of their transmission pattern.

From the zone-wise yearly data of dengue incidences from 2009 to 2013 and in 2015, the case vectors for each zone were formulated. These vectors were given as input to spherical *k*-means clustering algorithm. Package *sk*-means from R-software was used to perform clustering. From the elbow method, the number of clusters was determined. Figure 3 shows the diagram of the elbow method.

As observed from Figure 3, the angle is formed at the number of cluster 2. This implies that more than two clusters do not add much information as the sum of square error became steady after that. Here out of 12 zones in Delhi, only 9 zones were considered according to data availability.

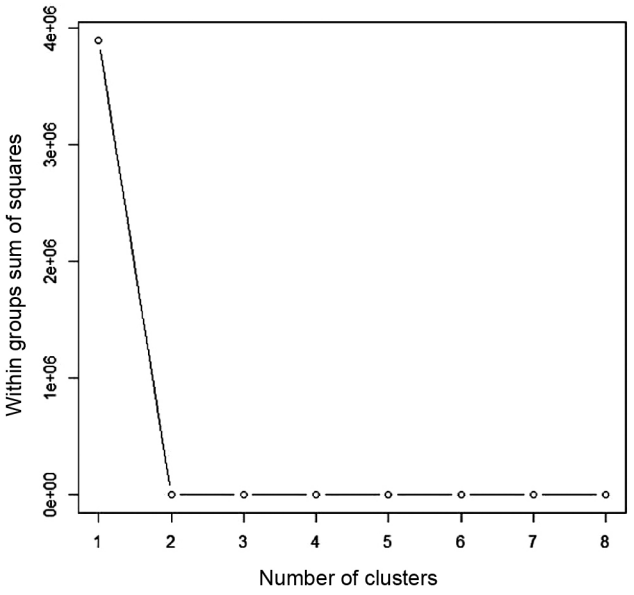


Figure 3. Result of elbow method.

Table 3. Adjusted *R*-squared values for various models

	Adjusted *R*-squared	
Degrees of freedom	Linear model	Quadratic model
3	0.64	0.64
6	0.21	0.27
9	0.49	0.60
12	0.54	0.62

Results

Multi-variate regression analysis

Multi-variate regression analysis has provided the relation between dengue and maximum temperature and rainfall. The models were retrospectively validated. The temporal data of 2006–2012 was given for training and 2013–2015 data were used for prediction and validation.

We took various degrees of freedom (3, 6, 9 and 12) for spline smoothing. Table 3 gives the adjusted R^2 values for various degrees of freedom. The models with degrees of freedom 3 and 9 failed in generalizing the model for test data. The model with degrees of freedom 6 gave good results, but the adjusted R^2 value for this model was very low. The model with 12 degrees of freedom gave good adjusted R^2 value and results. In Table 3 we notice that the model with degrees of freedom 3 has higher adjusted R^2 value. Figure 4 explains why the model with degrees of freedom 12 is better than this model. Figure 4 gives the residual plots of models with degrees of freedom 12 in Figure 4 *a* and degrees of freedom 3 in Figure 4 *b*. The perfect regression model fits a flat line through this scatter plot. We want this line (represented by a solid line) to be as flat as possible and it is clear that the model with degrees of freedom 12 achieves this condition better. Hence, it was chosen for spline smoothing the training data. Figure 5 shows the regression plot of linear and quadratic models.

As seen from Figure 5, the quadratic model can explain the peak in data better than the linear model. These peaks are necessary as they represent the outbreak. Also, the adjusted R^2 of linear and quadratic models are 0.54 and 0.62 respectively. We have chosen the quadratic model due to its higher adjusted R^2 and better results. The model equation is

$$\text{cases} = [-0.018 * (S(\text{temp}))^2] + [0.044 * (S(\text{rain}))^2] - 14.327, \quad (2)$$

where cases = dengue cases, temp = average of daily maximum temperature, rain = monthly total rainfall, S(temp) = cubic spline smoothing of temp, S(rain) = cubic spline smoothing of rain.

The graphs in Figure 6 compare the prediction result from model eq. (1) and actual results. These predictions are for entire Delhi. It is observed that although the predicted number of cases deviates from the actual number of cases, the trend followed in the cases is near the actual trend. For 2014, the results seem to be off-trend and over-predicted. Yet, it is clear that the predicted cases are

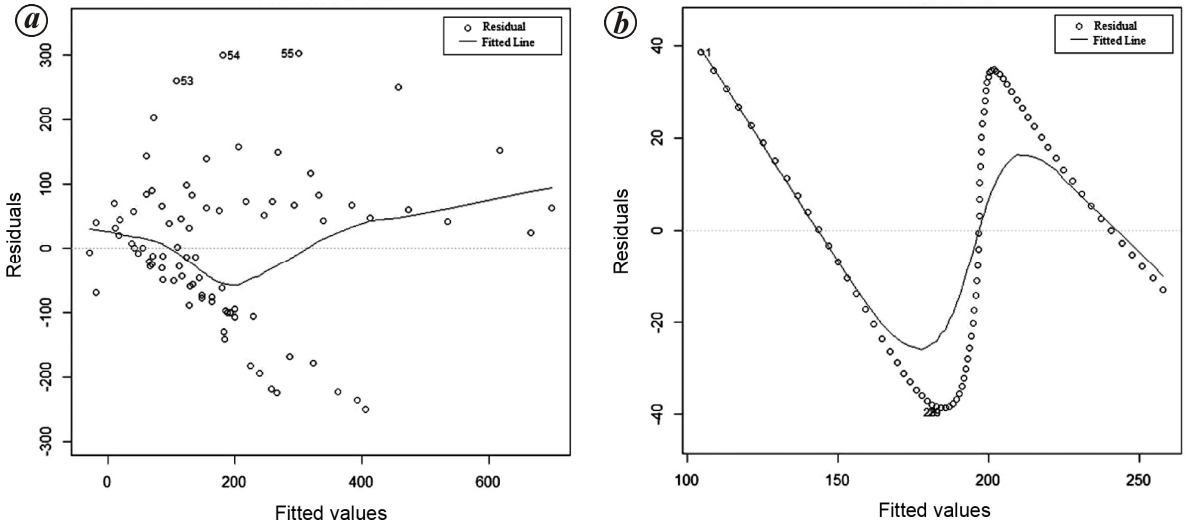


Figure 4. Residual plots of regression models: ***a***, model with degrees of freedom 12; ***b***, model with degrees of freedom 3.

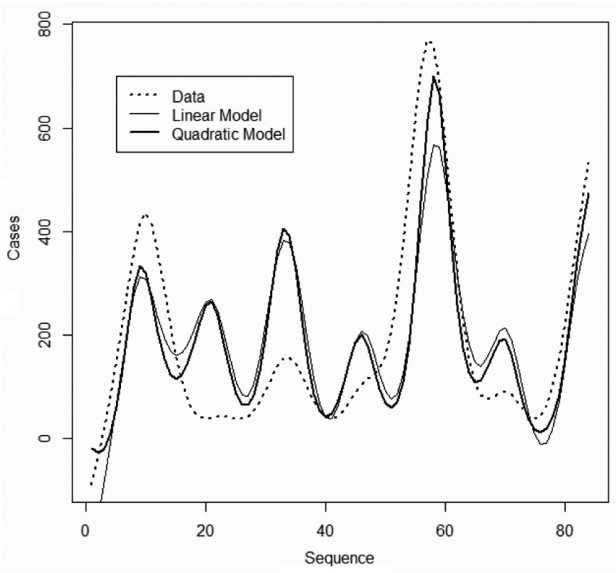


Figure 5. Regression plot in *R*.

Table 4. Fraction of cases for each zone

Zone	Fraction of cases
Central	0.095
Civil Line	0.069
Delhi City	0.023
Karolbagh	0.068
Najafgarh	0.093
Narela	0.033
Rohini	0.11
South	0.094
West	0.093

low when compared to the neighbouring years, 2013 and 2015. This shows that the model is successful in predicting a dip and rise in dengue cases over the years, which is crucial in predicting an outbreak.

Zone-wise prediction of dengue cases

As mentioned earlier, the modified model was used to predict the zone wise case of Delhi as well. For zone wise prediction of dengue cases the model was refined for local trend prediction as

$$\text{cases} = \text{cf} * [\{-0.018 * (S(\text{temp}))^2\} + \{0.044 * (S(\text{rain}))^2\} - 14.327], \quad (3)$$

where cf = fraction of cases for each zone, cases = dengue cases per zone, temp = zone wise average of daily maximum temperature, rain = zone wise monthly rainfall, S(temp) = cubic spline smoothing of temp and S(rain) = cubic spline smoothing of rain.

Table 4 gives the fraction of cases for each zone (value of cf in eq. (2)). Figure 7 shows comparison graphs for the predicted and actual dengue cases. The yearly trend for each zone can be easily compared from these graphs. Figure 8 shows the predicted and actual dengue heat maps for Delhi for 2011, 2012 and 2015 respectively. It must be noted here that the humidity parameter was also included in the regression study but the regression model failed in fitting the parameter and hence was excluded from the regression study.

Naïve Bayes

The Naïve Bayes model classified zones of Delhi into outbreak and non-outbreak zones using maximum temperature, relative humidity, rainfall and population density. Out of nine zones, for which data was available, six zones were used for training and three zones were used for validation for the years 2011, 2012 and 2015. Table 5 shows the classification results and Table 6 reports confusion matrix for Naïve Bayes results.

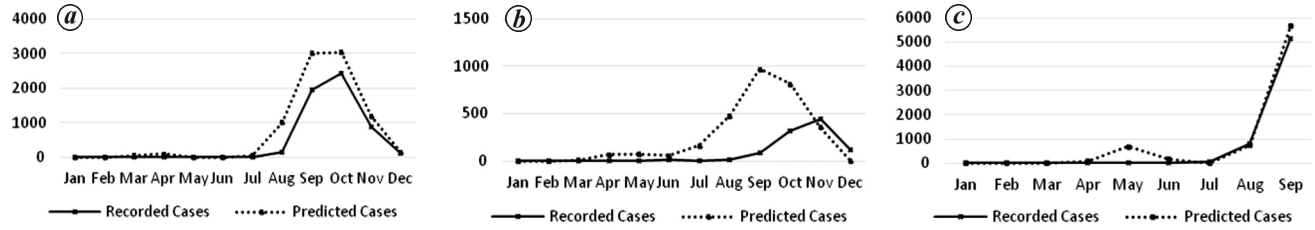


Figure 6. Prediction result for: ***a***, 2013; ***b***, 2014; ***c***, 2015.

Table 5. Naïve Bayes classification result

Year	Zone	Predicted Class	Actual Class
2011	Rohini	Non-outbreak	Non-outbreak
2011	South	Outbreak	Non-outbreak
2011	West	Non-outbreak	Non-outbreak
2012	Rohini	Non-outbreak	Non-outbreak
2012	South	Outbreak	Outbreak
2012	West	Non-outbreak	Non-outbreak
2015	Rohini	Outbreak	Outbreak
2015	South	Outbreak	Outbreak
2015	West	Outbreak	Outbreak

Table 6. Confusion matrix for Naïve Bayes results

		True class	
		Outbreak	Non-outbreak
Predicted class	Outbreak	4	1
	Non-outbreak	0	4

Table 7. Zones in each cluster

Cluster	Zones
Cluster 1	Narela, Delhi City, Najafgarh, Karolbagh
Cluster 2	Rohini, Civil Lines, Central, South, West

Spherical k-means cluster analysis

Figure 9 *a* shows the cluster analysis result. Here the numbers are indices of zones. Table 7 gives the distribution of zones into two clusters. The number of clusters is decided from the elbow method (Figure 3) and hence, two clusters are formed. This cluster analysis explains 73.61% of the total variance.

We have used the silhouette coefficient to evaluate the clusters. This is an internal evaluation method which uses the same data used for cluster analysis. The silhouette is defined as: $s(i) = b(i) - a(i)$ $\max\{a(i), b(i)\}$, where $s(i)$ = silhouette, $a(i)$ = average dissimilarity within the cluster for which i is member, $b(i)$ = lowest average dissimilarity between clusters where i is not a member. The value of $s(i)$ is in the range [–1, 1]. For a perfect cluster this value will be exactly 1. The negative value indicates that the object in question would be more suitable in another cluster. Figure 9 *b* shows the silhouette values for each cluster.

Discussion

Our study implemented two methods for dengue outbreak detection using weather parameters. Both methods have shown that there is a strong influence of weather parameters in shaping dengue cases. Results have identified the relation between dengue and weather parameters. Regression analysis successfully modelled the effects of maximum temperature and rainfall. Initially, through regression analysis, the significance of relative humidity was not identified. Application of Naïve Bayes was useful in overcoming this limitation. Naïve Bayes incorporated effects of relative humidity and population density. Figure 10 shows the variation in these weather parameters with variation in dengue cases. All these graphs have a lag period of 2 months. As seen in Figure 10 *a* and *b*, the dip in maximum temperature and relative humidity and peak in dengue cases coincide. Dengue cases are reported during and after monsoon season – August to November. During this time maximum temperature drops to around 20–25°C. At temperatures below 14–15°C survival of mosquitoes and their eggs becomes difficult[28]. Hence, as temperature continues to drop, there is a decline in dengue cases towards the end of November. As monsoon starts, relative humidity starts increasing which had dropped to 50–60% during summer. Similarly, from Figure 10 *c*, it is clear that a strong pattern visible between dengue cases and rainfall. The years 2006, 2011, 2013 and 2015 were the dengue outbreak years. In the graph, it is visible that these years had a sudden increase in rainfall (>300 mm), which might have facilitated mosquito breeding resulting in a higher number of cases.

The relation of dengue with weather variables represents the effects of weather parameters in shaping dengue cases which is our main objective. As pointed out[29], high rainfall and temperature lead to high vapour pressure, a condition favourable for breeding and survival of mosquitoes. To control dengue, vector control measures become obligatory as no vaccine is available yet for dengue and poor weather conditions make this a difficult process[29].

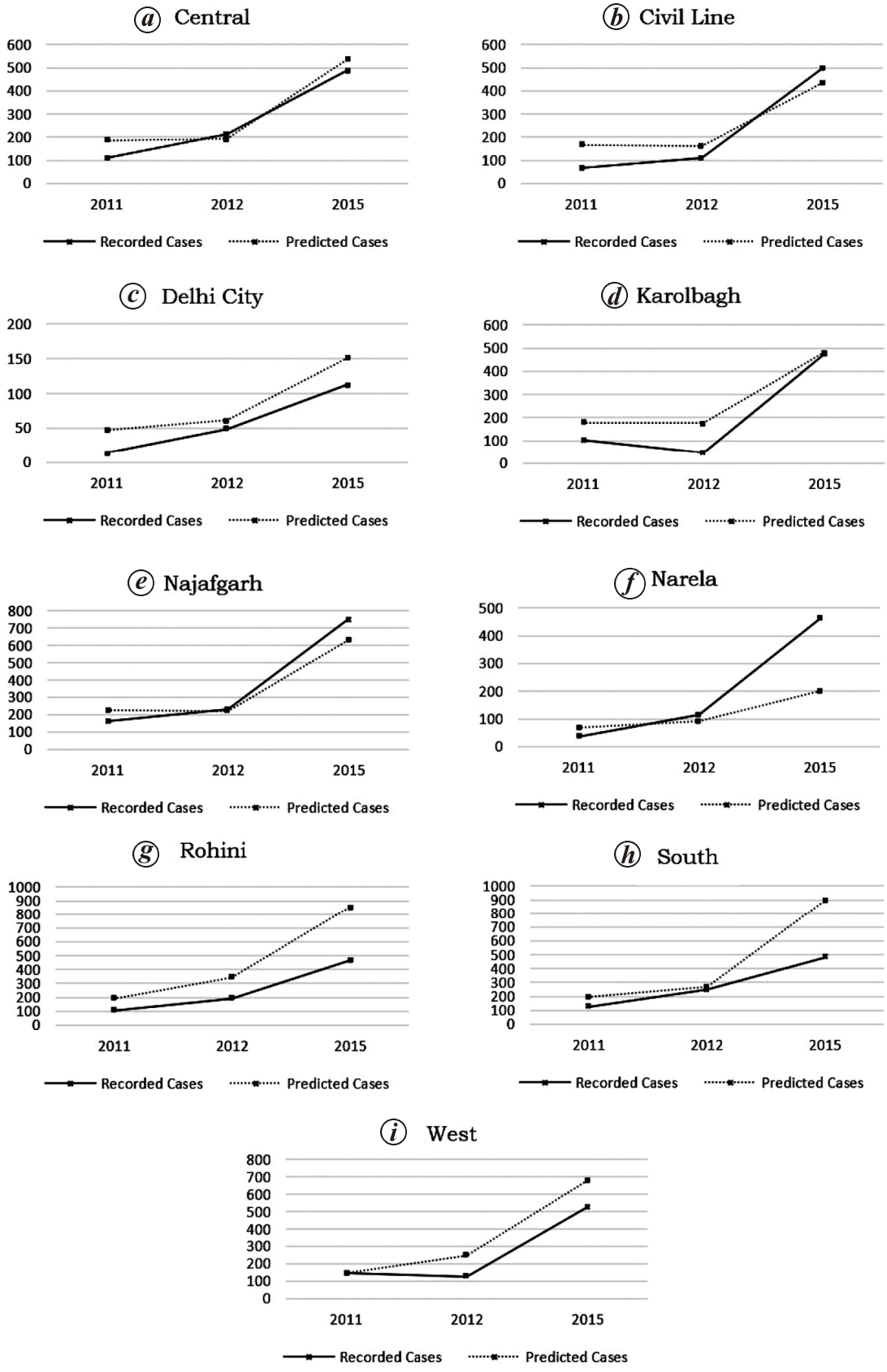


Figure 7. Zone wise prediction for Delhi. ***a***, Central; ***b***, Civil line; ***c***, Delhi city; ***d***, Karolbagh; ***e***, Najafgarh; ***f***, Narela; ***g***, Rohini; ***h***, South; ***i***, West.

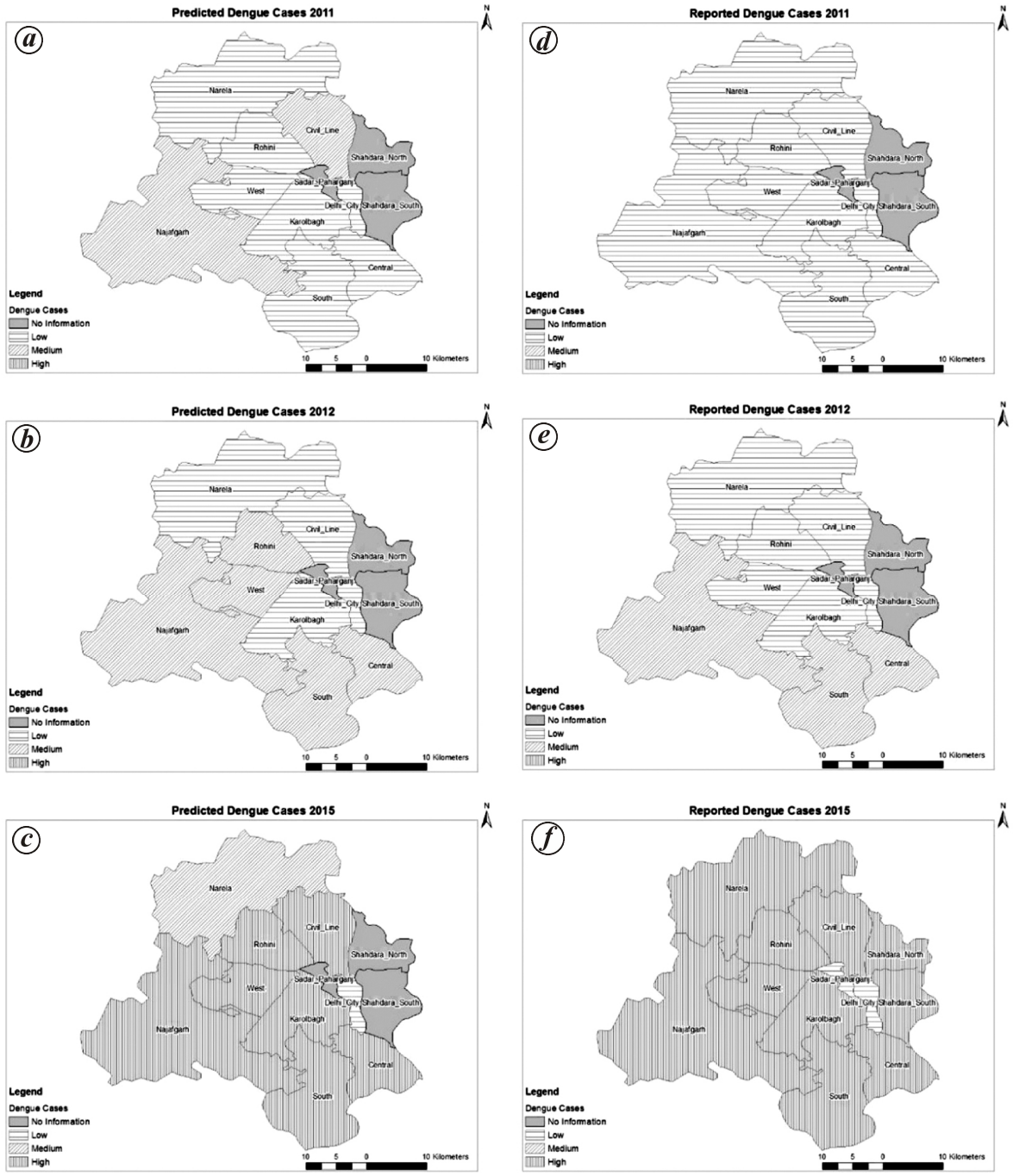


Figure 8. Dengue heatmaps for Delhi. These maps are the prediction result from our model and recorded cases in Delhi in the following order: ***a***, Dengue prediction for 2011; ***b***, Dengue prediction for 2012; ***c***, Dengue prediction for 2015; ***d***, Dengue recorded cases for 2011; ***e***, Dengue recorded cases for 2012; ***f***, Dengue recorded cases for 2015.

This modelled relation will, however, allow estimation of these effects by weather parameters to some extent. These relations indicate that sudden and high rainfall with total rainfall >300 mm accompanied with a maximum temperature of around 30–35°C and high relative humidity are ideal conditions for dengue cases to spread. When these conditions are satisfied, there is a high probability of a large number of reported dengue cases, and if no precautions are taken, then the spread can result in an outbreak.

Apart from this, cluster analysis has grouped Delhi zones into transmission zones. Figure 11 shows the map of transmission zones clustered using spherical k-means algorithm. The cluster 2 is significant because the zones

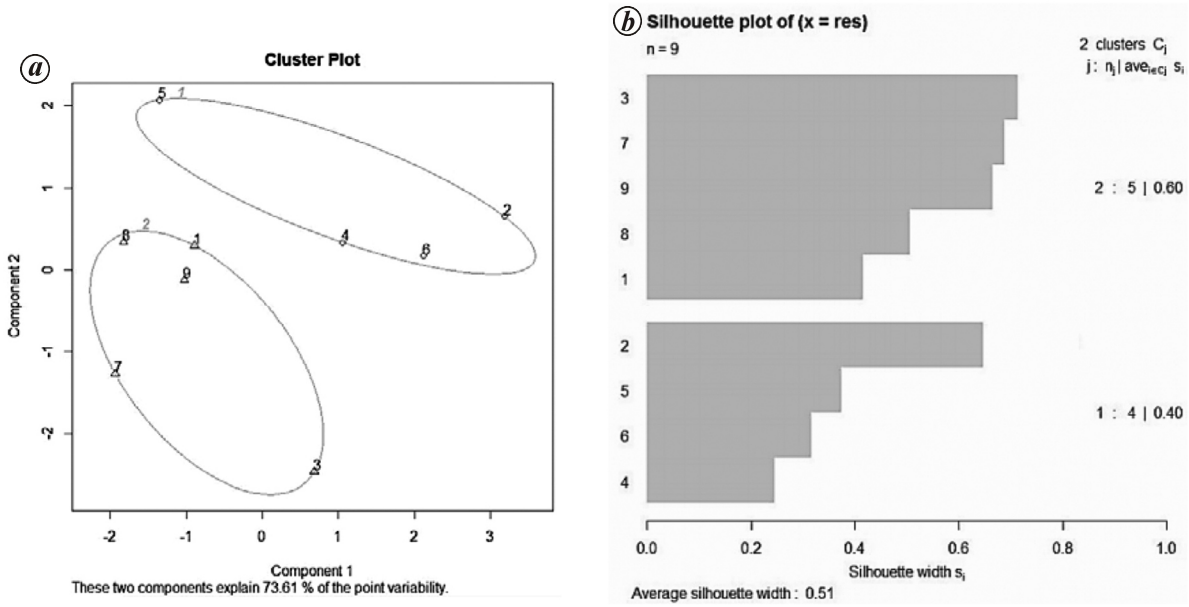


Figure 9. ***a***, Cluster plot of spherical *k*-means. ***b***, Silhoutte values of clusters.

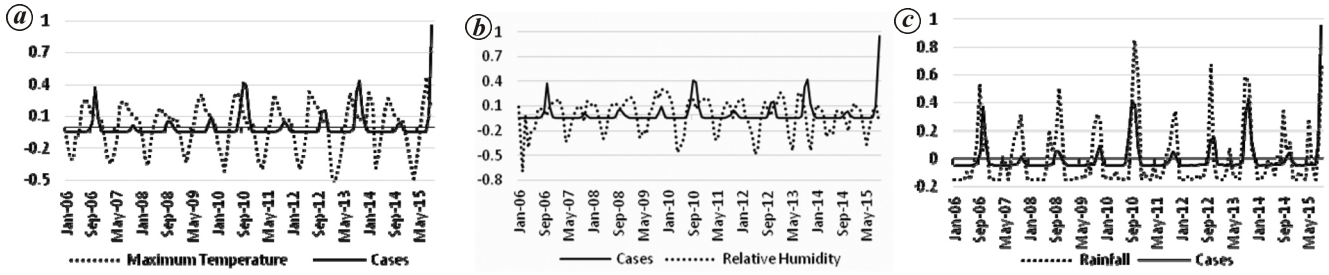


Figure 10. ***a***, Maximum temperature and dengue cases. ***b***, Relative humidity and dengue cases. ***c***, Rainfall and dengue cases.

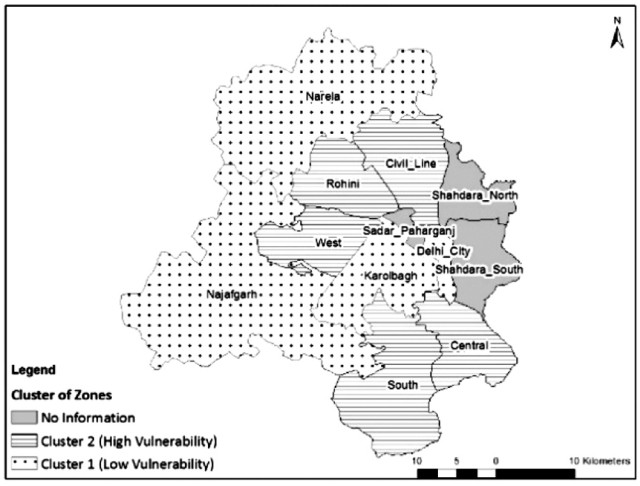


Figure 11. Transmission zones.

in cluster 2 report a high number of dengue cases. As these are neighboring zones, movement of infected people between these zones can be one of the reasons for high vulnerability in these zones. These zones are prone to the spread of dengue whenever weather conditions facilitate mosquito breeding and hence require high precaution during monsoon season. Cluster 1, on the other hand, is less vulnerable according to historical data.

Conclusions

Our study describes the relationship between weather parameters and dengue incidences. We have used regression and Naïve Bayes to identify this relationship. Here, the advantage of Naïve Bayes over regression analysis is that it was able to identify the significance of relative humidity and population density. Another advantage is that Naïve Bayes can easily adapt to new data because it incorporates the effects of population as well as the area in the form of population density. The regression model, on the other hand, is highly dependent on the training data and different relationships for different regions can exist. One disadvantage of Naïve Bayes is that it works well only with a small dataset. Considering our application

and data availability either approach can be used to model the relationship. Using this relationship, an early warning system can be developed which allows decision makers to take important decisions to prevent and control dengue outbreaks. Also, cluster analysis showed the zones which are similar in terms of the dengue transmission pattern. This analysis grouped highly vulnerable zones into one cluster and others in another cluster. From these results transmission zone map was prepared which can further be used to analyse the high reporting of dengue incidences in certain zones.

Data scarcity was the biggest challenge of this study. The zone-wise aggregated cases were available but their geographical distribution in the form of coordinates was not available and therefore our study was limited to few zones. Other parameters like immunity, the serotype of dengue, demographics are also substantial which were not modelled due to unavailability of data. If these parameters are also incorporated in the study, it would provide better prediction and help remove the deviations currently present.

1. Bhatt, S. *et al.*, The global distribution and burden of dengue. *Nature*, 2013, **496**, 504–507.
2. WHO, Dengue: Guidelines for diagnosis, treatment, prevention and control, World Health Organization, 2009.
3. Powell, J. R. and Tabachnick, W. J., History of domestication and spread of *Aedes aegypti* – a review. *Mem. Inst. Oswaldo Cruz*, 2013, **108**, 11–17.
4. Gubler, D. J., Dengue and dengue hemorrhagic fever. *Clin. Microbiol. Rev.*, 1998, **11**, 480–496.
5. Gupta, N., Srivastava, S., Jain, A. and Chaturvedi, U. C., Dengue in India. *Indian J. Med. Res.*, 2012, **136**, 373–390.
6. Shepard, D. S. *et al.*, Economic and disease burden of dengue illness in india. *Am. J. Trop. Med. Hyg.*, 2014, **91**, 1235–1242.
7. Rao, T. R., Distribution, density and seasonal prevalence of *Aedes aegypti* in the indian subcontinent and south-east asia. *Bull. World Health Organ.*, 1967, **36**, 547.
8. Nimmannitya, S., Gubler, D., Biswas, A., Devgan, V., Gupta, B. and Sharma, S., Guidelines for clinical management of dengue fever, dengue haemorrhagic fever and dengue shock syndrome. Programme, D.o.N.V.B.D.C.).
9. Cecilia, D., Current status of dengue and chikungunya in India. *WHO South-East Asia J. Public Health*, 2014, **3**, 22–27.
10. Vikram, K. *et al.*, An epidemiological study of dengue in Delhi, India. *Acta Trop.*, 2015, **153**, 21–27.
11. Kukreti, H. *et al.*, Emergence of an independent lineage of dengue virus type 1 (denv-1) and its co-circulation with predominant denv-3 during the 2006 dengue fever outbreak in Delhi. *Int. J. Infect. Dis.*, 2008, **12**, 542–549.
12. Chan, M. and Johansson, M. A., The incubation periods of dengue viruses. *PLoS ONE*, 2012, **7**, e50972.
13. Carrington, L. B., Armijos, M. V., Lambrechts, L. and Scott, T. W., Fluctuations at a low mean temperature accelerate dengue virus transmission by *Aedes aegypti*, 2013.
14. Lambrechts, L., Paaijmans, K. P., Fansiri, T., Carrington, L.B., Kramer, L. D., Thomas, M. B. and Scott, T. W., Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *Proc. Natl. Acad. Sci. USA*, 2011, **108**(18), 7460–7465.
15. Brady, O. J. *et al*, Global temperature constraints on *Aedes aegypti* and ae. Albopictus persistence and competence for dengue virus transmission. *Parasit. Vectors*, 2014, **7**, 1–4.
16. Nakhapakorn, K. and Tripathi, N. K., An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *Int. J. Health Geogr.*, 2005, **4**, 13.
17. Hii, Y. L., Rocklöv, J., Ng, N., Tang, C. S., Pang, F. Y. and Sauerborn, R., Climate variability and increase in intensity and magnitude of dengue incidence in singapore. *Glob. Health Action*, 2009, **2**(1), 2036.
18. Promprou, S., Jaroensutasinee, M. and Jaroensutasinee, K., Climatic factors affecting dengue haemorrhagic fever incidence in southern Thailand. *Dengue Bull.*, 2005, **29**, 41.
19. Karim, M., Munshi, S. U., Anwar, N. and Alam, M., Climatic factors influencing dengue cases in Dhaka city: a model for dengue prediction. *Indian J. Med. Res.*, 2012, **136**, 32.
20. Fathima, A. S., Manimegalai, D. and Hundewale, N., A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. *Int. J. Comput. Sci.*, 2011, **8**(6).
21. Bakar, A. A., Kefli, Z., Abdullah, S. and Sahani, M., Predictive models for dengue outbreak using multiple rulebase classifiers. In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on IEEE, pp. 1–6.
22. Shaukat, K., Masood, N., Mehreen, S. and Azmeen, U., Dengue fever prediction: A data mining problem. *J. Data Min. Genom. Proteomics*, 2015, **6**(181), 2153–0602.
23. Shakil, K. A., Anis, S. and Alam, M., Dengue disease prediction using weka data mining tool. arXiv preprint arXiv:150205167, 2015.
24. Prakash, M. and Kumar, S. C., Hotspot analysis of dengue fever cases in delhi using geospatial techniques. In *2014 Esri India User Conference*, New Delhi, India.
25. Lewis, D. D., Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: Ecml-98*, Springer, pp. 4–15.
26. Kumar, S. A. and Vijayalakshmi, M., Inference of naïve baye's technique on student assessment data. In *Global Trends in Information Systems and Software Applications*, Springer, pp. 186–191.
27. Hornik, K., Feinerer, I., Kober, M. and Buchta, C., Spherical k-means clustering. *J. Stat. Softw.*, 2012, **50**, 1–22.
28. Brady, O. J. *et al.*, Modelling adult aedes aegypti and aedes albopictus survival at different temperatures in laboratory and field settings. *Parasit. Vectors*, 2013, **6**, 1–12.
29. Hales, S., De Wet, N., Maindonald, J. and Woodward, A., Potential effect of population and climate changes on global distribution of dengue fever: An empirical model. *Lancet*, 2002, **360**, 830–834.

Received 23 July 2016; revised accepted 11 February 2018

doi: 10.18520/cs/v114/i11/2281-2291