Journal of Rural Development, Vol. 37, No. (2), pp. 245-258
NIRD&PR, Hyderabad.

SPATIAL MODELLING OF INTERACTIONS BETWEEN DENGUE INCIDENCES AND CHANGING CLIMATE BY INTEGRATING ANN TECHNIQUE WITH GIS

Shuchi Mala
and
*Mahesh Kumar Jat**

ABSTRACT

According to various studies, it is well established that climate characteristics are one of the significant factors influencing vector-borne diseases and their long-term variations in a climate change scenario may affect the vector-borne diseases. According to World Health Organisation (WHO) factsheets (2016), more than 2.5 billion people in over 100 countries are at risk of dengue alone which is one of the deadliest vector-borne diseases. Therefore, it is very vital to develop a surveillance system which is capable of predicting the high-risk areas so that the proactive and effective control measures can be taken immediately. In the present study, a spatial data mining model is developed by integrating artificial neural network (ANN) technique into a Geographic Information System (GIS). A statistical method such as logistic regression has been used to detect the areas where the prevalence of the disease is high. Also, possible associations between disease incidences and meteorological parameters have been investigated. This model will highlight areas which are at high risk of dengue by examining the interactions between dengue fever incidences and environment. The primary purpose of the present work is to provide a better understanding of the spatial dispersal of the dengue fever risk in the rural as well as the urban areas of Delhi. Also, this model will bring new insights to the public health officials and policymakers to reduce the risk of deaths mainly in rural areas due to lack of awareness and health facilities.

Keywords: Dengue, Outbreaks, Risk, Artificial Neural Networks, Regression.

* Research Scholar and Associate Professor, Respectively, Civil Engineering Department, Malaviya National Institute of Technology, Jaipur-302017. INDIA, E-mail:2013rce9514@mnit.ac.in

Authors are thankful to Dr. Ashok Rawat, Deputy Health Officer, Karol Bagh Zone, and Dr. Sanjay Sinha, Deputy Health Officer, Rohini Zone, Health Department, Municipal Corporation of Delhi for providing the data on Dengue Fever incidences to perform the investigation. Authors are thankful to Fund for Improvement of S&T Infrastructure (FIST) of the Department of Science & Technology (DST), Government of India for funding the study. The authors are thankful to Ministry of Human Resource Development, India for providing fellowship for carrying out the research work.

Introduction

The vector-borne diseases such as dengue is one of the most rapidly spreading mosquito-borne diseases. According to World Health Organisation (WHO), in the world, the estimated number of dengue occurrences are 50-100 million every year of which 15-30 million incidences happen in India (WHO factsheet 2013, 2014). Prevention and control of dengue fever (DF) is an effective method to reduce the mortality rate. GIS plays a vital role in disease susceptibility mapping to control the spread of the disease. The generated risk maps help to identify potential areas of DF due to their social and environmental conditions.These maps predict the distribution of DF incidences which help public health officials and planners to develop effective control strategies and thus enabling the establishment of early warning system that could contribute in minimisation of the DF occurrences due to the influence of various causative factors. According to various studies, meteorological parameters have a significant impact on DF incidences.This impact is due to parameters such as temperature and humidity which affect the life cycle, the rate at which mosquito bites, infectious and survival rates of mosquitoes and on the incubation period of dengue virus (Hii et al., 2009). As temperature increases, Aedes mosquitos',carrier of dengue virus displays shorter periods of growth in all stages of life-cycle causing an increase in the vector density.Xu et al. (2014) investigated the influence of meteorological parameters on dengue in Singapore.Two models such as Poisson regression and distributed Lag Non-linear Model have been combined to evaluate the impact of mean, minimum and maximum temperatures, rainfall, absolute humidity, relative humidity and wind speed on dengue incidences (2001-2009). Mean temperature, absolute humidity and rainfall impact on dengue incidences were found significant. Positive association of dengue cases was observed with temperature and absolute humidity.Ughelli et al. (2017) performed an analysis of the influence of climatic variables on DF in Paraguay. The efficiency of the neural networks is used to predict the number of disease cases. A variable selection and prediction method that can be used for any geographical region was developed showing favourable results.Therefore, it is evident that the meteorological parameters are influencing the spatial distribution of DF incidences.

According to various studies, susceptibility maps could be generated by modelling the relationship between DF incidences and essential physiographic and environmental factors (Tsegaw et al., 2013; Yang, 2016). There are two types of disease mapping techniques such as knowledge-driven and data-driven.The knowledge-driven approach is based on the expert's knowledge to assign weights to a set of factors whereas data-driven approach is based on the retrospective data patterns and relationships. The advantage of data-driven approach over knowledge-driven approach is that it provides a more straightforward and more direct method for disease susceptibility mapping. Among the various data-driven methods, logistic regression (LR) and artificial neural networks (ANNs) are the two most widely used methods. LR is a multivariate regression technique that is

used to predict the probability of presence or absence of a particular condition. This statistical method has been extensively used to generate risk maps. ANNs are widely used to manage multi-dimensional non-linear features of practical problems. However, very few studies have used these techniques in producing disease susceptibility maps.

In the present study logistic regression has been used to determine the association between meteorological parameters and DF incidences. Also, an attempt has been made to implement radical basis functional network for producing disease susceptibility maps using Python and ArcGIS.

Methodology

In the present study, the methodology can be classified into two sections. Firstly, identification of an association between critical meteorological parameters and DF incidences. Secondly, if the relationship exists then, ANN is used to identify the risk-prone areas of DF outbreak in the study area. ANN epitomises as a powerful data-driven method that models the behaviour of DF incidences by investigating their association with meteorological parameters in the study area. The proposed model is aimed to perform predictive modelling of potential areas of DF incidences.

Study Area: In the present retrospective study, DF incidences data have been obtained for Delhi and the study period is 2006-2013. Delhi covers an area of 1,484 km2. It has a length of 51.9 km and a width of 48.48 km (Vikram et al., 2015). It is located at 28.7041° N latitude and 77.1025° E longitude. It has a population of about 16.3 million (Census 2011) making it the second most populous city in India. These comprise two civic bodies, the cantonment authority and the Railways. The Municipal Corporation of Delhi (MCD) area is divided into 12 zones spread through three smaller municipal corporations and 272 wards covering about 1399 km2 area. The New Delhi Municipal Council (NDMC) has one zone and eight wards covering 42.74 km2 area. The study area Delhi has been shown in Figure 1.

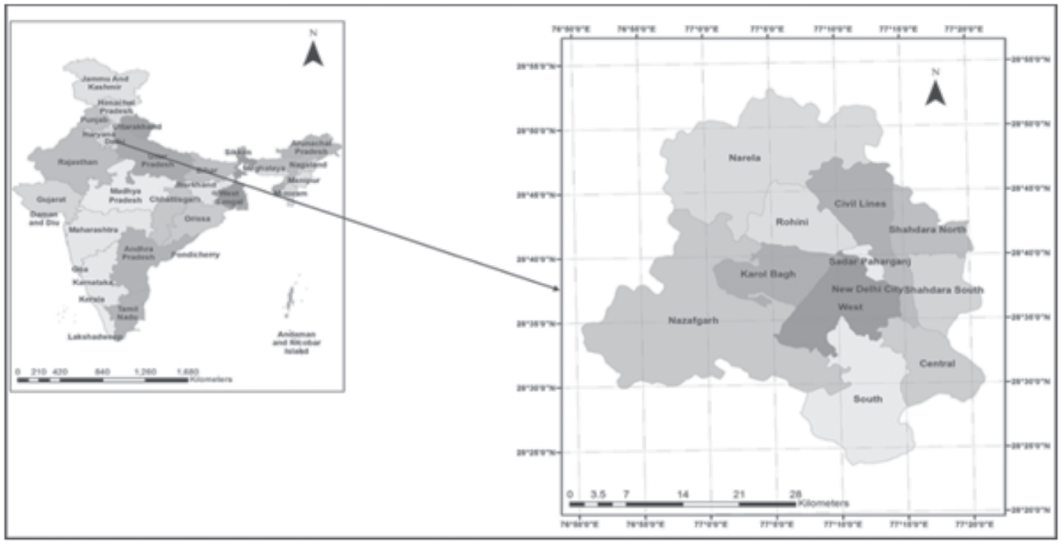

Figure 1: Map for the Study Area-Delhi

Data Resources: In the present study, data collection include survey of India toposheets, meteorological data and infectious disease incidence data. Meteorological data have been collected from India Meteorological Department (IMD) for Delhi and the study period is 2006 to 2013. Data have been collected from 13 weather stations (Palam, Safdarjung, Ayannagar, Sports Complex, NCMRWF, Pusa, Pitampura, Mungeshpur, Jafarpur, Delhi University, DPS Indirapuram, Narela and Akshardham) at Delhi. The data have been collected at hourly scale for meteorological parameters such as minimum temperature (°C), relative humidity (%) and mean wind speed (km/h). Daily cases of DF from the year 2006 to the year 2013 have been taken for the study area from Health Department of Municipal Corporation of Delhi (MCD). MCD is a government agency of India which brings together confirmed cases of epidemics collected from different hospitals. The data included the date, residential location, age and gender of the notified cases of DF.

Logistic Regression: Regression can be defined as a process to determine a series of coefficients that describe the association between the independent variables and the dependent variables effectively. LR is a regression technique used to determine the relationship between several independent variables and the probability of a binary or categorical response (Lee and Sambath, 2006). The main advantage of LR is that when a suitable link function is added to a linear regression model then the variables can be any combination of continuous and discrete variables and they do not have to follow a normal distribution (Lee and Sambath, 2006). This regression technique could be used to predict the probability of disease incidences by considering various critical causative factors because the value of response variable is the

probability of the disease incidences.The association of a dependent variable with independent variables is mathematically expressed in equation 1 and equation 2.

$$P = 1/_{(1+e-z)} \quad (1)$$

$$Z = b_o + b_1 x_1 + b_2 x_2 + ... + b_n x_n \quad (2)$$

In equation 1, P is the probability of the occurrence of an event and Z is the linear combination of the independent variables. In the present study, Z is a linear combination of meteorological factors such as minimum temperature, absolute humidity and mean wind speed. In equation 2, b_o is the intercept of the LR model and $b_1, b_2, \ldots b_n$ are the weights of the factors.

Radial Basis Functional Network: ANN belongs to the family of machine learning and cognitive science, inspired by the sophisticated functionality of human brain which is responsible for processing information coming from hundreds of billions of inter-connected neurons in parallel.The typical structure of ANN comprises three layers such as input layer, a hidden layer with several neurons and output layer, with each layer entirely linked to the succeeding one through a sequence of weights.

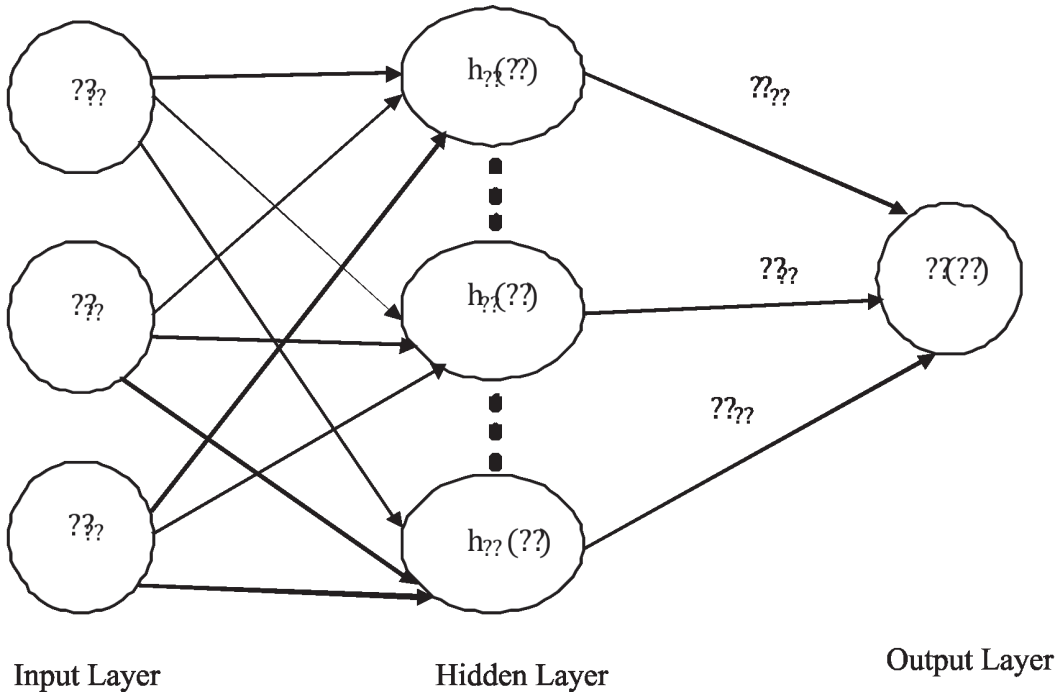

Figure 2: Basic Structure of RBFN

The neurons having transfer functions can obtain signals approaching from the preceding layer and produce outputs which form the input signals for the succeeding layer.The adjustment of weights is based on the empirical data which make the neural nets having the capability of learning.In the present study, RBFN type of ANN has been used to produce disease susceptibility maps.RBFN are used to determine the non-linear input-output relationships. It comprises of three layers such as 1) an input layer with n nodes that receive values from observed data; 2) a hidden layer which consists of m artificial neurons; 3) an output layer.The active function in the neurons in the hidden layer is the radial basis function (RBF).

$$y=f(x,v)=\exp\left[-||x-v||^2/(2s\sigma 2)\right] \quad (3)$$

The equation 3 is showing the common Gaussian RBF where the x presents an n-dimensional vector, v represents the centre, and s is the spread variable. The purpose of spread parameter is the determination of the size of the receptive field.

$$\sigma = (1/4)(1/M)1/N \quad (4)$$

The equation 4 is showing the mathematical formulation of spread parameter. In the equation, M is the number of hidden neurons and N the number of the predictor maps.

$$f(X) = \sum \frac{m}{j=1} wjhj^{(X)} \quad (5)$$

$$h(x) = \exp\left(-\frac{(x-c)^2}{r^2} \right) \quad (6)$$

In the equations 5 and 6, h(X) is the Gaussian activation function with the parameter r. The parameter r is the radius or standard deviation, and c is the centre or average taken from the input space.The parameter c is defined separately at each RBF unit.The learning process is based on adjusting the parameters of the network to reproduce a set of input-output patterns. When RBFN is initialised, one training data received by the input layer is associated with all the neurons in the hidden layer.The observed data which are transmitted to the hidden layer are fed into the radial basis function to produce an output. Then the output produced by the hidden layer and the weight-w (w_i,w_j,…,w_m) form a dot product which is transmitted to the output layer. Lastly, the output layer responses the output of the fed data. The desired weight w of the RBFN could be trained using gradient descent approach (Looney, 2002).

Results and Discussion

In the present study, multi-nomial logistic regression has been performed.The independent variables are minimum temperature, absolute humidity and mean wind speed.The dependent variable is the category of risk of DF incidences. The number of DF incidences in each ward of MCD has been categorised into four levels. The first level is no risk, second is low risk, third is high risk, and the fourth level is very high risk. The study period is from 2006 to 2013. At each ward of Delhi and for the date of each reported case, meteorological parameters value have been extracted using interpolation technique that is inverse distance weighted (IDW). Absolute

humidity has been calculated using the temperature and relative humidity. Figure 3 is showing the distribution of DF incidences with respect to absolute humidity. The range of absolute humidity at which high number of DF incidences can be observed is 14 to 18(g)/m^3 . Figure 4 is showing the distribution of DF incidences with respect to minimum temperature. The range of minimum temperature at which a high number of DF incidences can be observed is 18 to 24°C. Figure 5 is showing the distribution of DF incidences with respect to mean wind speed.The range of mean wind speed at which a high number of DF incidences can be observed is 2 to 6 km/h.The distribution of DF incidences according to the meteorological parameters revealed that at a particular range of these parameters DF outbreaks occur.Multinomial logistic regression has been performed for three scenarios.The first scenario (Scenario-A) is when independent variables are minimum temperature and absolute humidity. The second scenario (Scenario-B) is when independent variables are absolute humidity and mean wind speed.The last scenario (Scenario-C) is when independent variables are minimum temperature and mean wind speed.

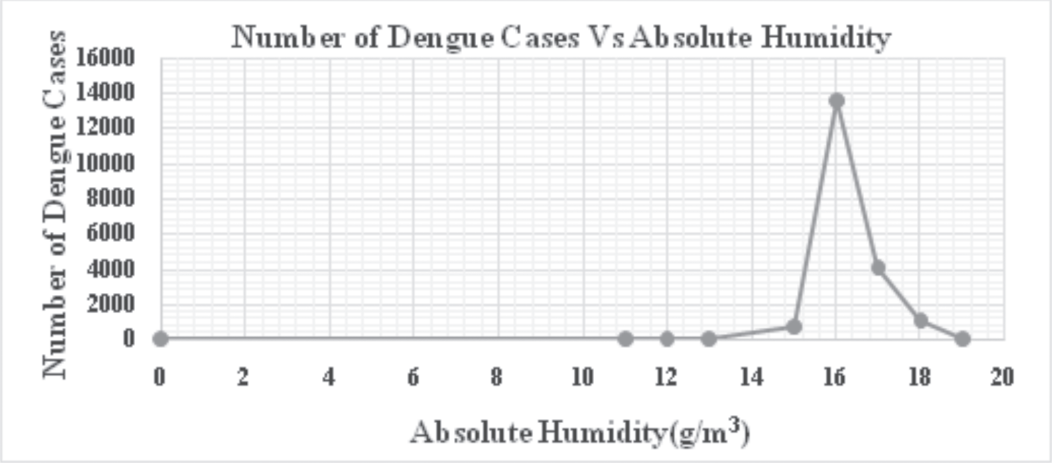

Figure 3: Distribution of DF Incidences with Respect to Absolute Humidity

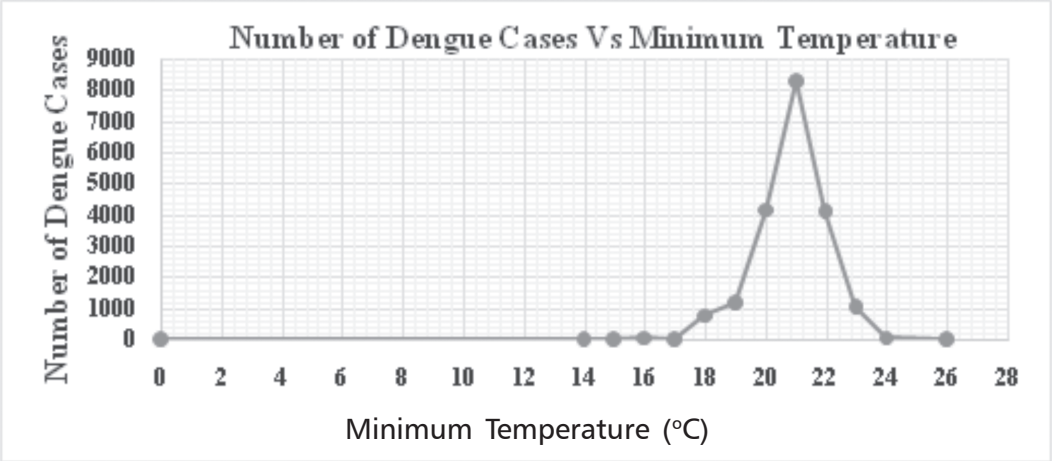

Figure 4: Distribution of DF Incidences with Respect to Minimum Temperature

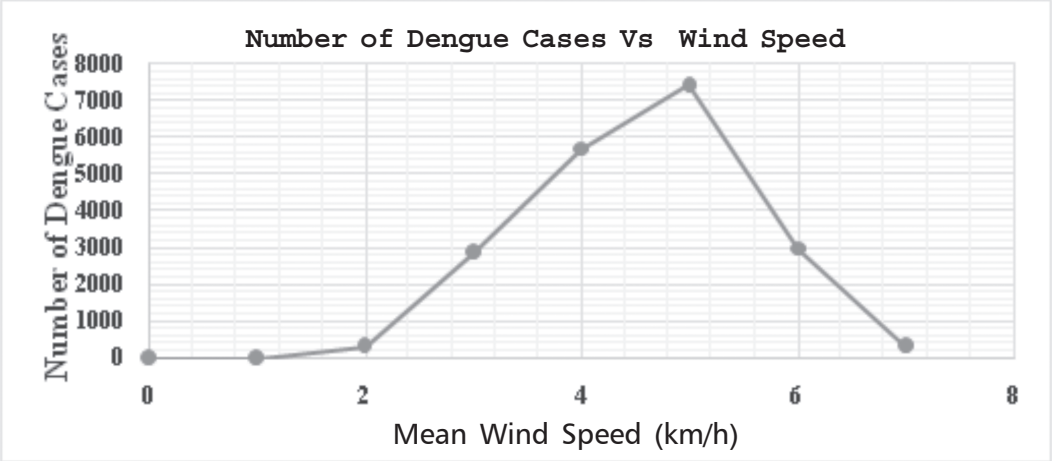

Figure 5: Distribution of DF Incidences with Respect to Mean Wind Speed

For each scenario (A-C), regression has been performed, and various statistics have been calculated such as R-squared value (R^2) which determines whether model fits the data well and standard errors which are the errors in the regression coefficients.Other statistics are Wald statistic used to determine whether explanatory variables are significant in a model or not and Pr>chi-squared which is defined as the probability of observing a Chi-Square statistic as extreme as or more so, than the observed one under the null hypothesis. Also, Akaike Information Criteria (AIC) has been calculated which is an estimator of the relative quality of statistical models for a given set of data. For each scenario regression equation has been

established. In Table 1, regression equation for each category of DF occurrences such as 2 (low risk), 3 (high risk) and 4 (very high risk) have been shown for each scenario. For Scenario-A, relationship between minimum temperature (X1), absolute humidity (X2) and level of DF occurrence risk has been mathematically formulated.This model has very high R-squared value and low AIC value (R^2 = 0.6, AIC = 79) thus showing very strong association between the variables. For Scenario-B (Table 1), relationship between absolute humidity (X1), mean wind speed (X2) and level of DF occurrence risk has been mathematically formulated.This model has high R-squared value and low AIC value (R^2 = 0.4, AIC = 85) thus showing strong association between the variables. Similarly, for Scenario-C (Table 1), relationship between minimum temperature (X1), mean wind speed (X2) and level of DF occurrence risk has been mathematically formulated.This model has low R-squared value and high AIC value (R^2 = 0.2, AIC = 93) compared to other two scenarios thus showing weak association between the variables.Other statistical parameters value have been shown in Table 2. Higher the Wald statistics, higher the accuracy of the model prediction.The Wald statistic is varying from 0 to 6. Highest Wald statistic is observed for Scenario-A.

Table 1: Multinomial Logistic Regression Model Parameters for Each Scenario

	Regression Equation		
Scenario	logit (y=2)	logit (y=3)	logit (y=4)
A	-0.31-1.966X1+2.655X2	-1.67-1.59X1+2.22X2	-5.151-2.08X1+3.05X2
B	-3.27+0.37X1-0.33X2	-4.53+0.45X1-0.46X2	-16.19+0.65X1+1.22X2
C	0.45+0.04X1-0.12X2	-1.47+0.15X1-0.32X2	-6.31+0.11X1+0.85X2

Thus, the results have shown that strong association exists between meteorological parameters such as minimum temperature and absolute humidity and level of DF occurrence risk, that is, Scenario-A model fits the data well as compared to other two scenarios. Therefore, these two meteorological parameters have been used to perform DF susceptibility mapping using ANN.

Table 2: Multinomial Logistic Regression Model Parameters for Each Scenario

Scenario	Level of DF risk	Source	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower bound (95%)	Wald Upper bound (95%)
A	2	Intercept	4.49	0.00	0.94	-9.12	8.49
		X1	0.93	4.51	0.03	-3.78	-0.15
		X2	1.28	4.33	0.04	0.15	5.16
	3	Intercept	4.82	0.12	0.73	-11.11	7.77
		X1	0.93	2.90	0.09	-3.41	0.24
		X2	1.29	2.94	0.09	-0.32	4.76
	4	Intercept	5.99	0.74	0.39	-16.90	6.60
		X1	0.95	4.73	0.03	-3.95	-0.21
		X2	1.36	5.04	0.02	0.39	5.70
B	2	Intercept	3.11	1.10	0.29	-9.37	2.83
		X1	0.20	3.43	0.06	-0.02	0.77
		X2	0.50	0.42	0.52	-1.31	0.66
	3	Intercept	3.99	1.29	0.26	-12.36	3.29
		X1	0.25	3.13	0.08	-0.05	0.94
		X2	0.62	0.54	0.46	-1.68	0.76
	4	Intercept	7.94	4.16	0.04	-31.76	-0.63
		X1	0.31	4.57	0.03	0.05	1.25
		X2	0.87	1.94	0.16	-0.49	2.93
C	2	Intercept	3.23	0.02	0.89	-5.89	6.79
		X1	0.17	0.06	0.81	-0.29	0.38
		X2	0.51	0.06	0.81	-1.12	0.88
	3	Intercept	3.82	0.15	0.70	-8.96	6.02
		X1	0.19	0.56	0.45	-0.24	0.53
		X2	0.61	0.28	0.60	-1.53	0.88
	4	Intercept	5.45	1.34	0.25	-16.98	4.36
		X1	0.22	0.23	0.63	-0.32	0.53
		X2	0.67	1.60	0.21	-0.47	2.16

RBFN has been used to produce dengue susceptibility maps.The type of RBFN used in the present study is a probabilistic neural network (PNN). Python programming has been used to develop a model which uses RBFN to perform classification and prediction of the level of DF occurrence risk.The training data consist of 2 per cent of no risk of DF occurrences data points,

38 per cent of the low risk of DF occurrences data points, 56 per cent of the high risk of DF occurrences data points and 4 per cent of the very high risk of DF occurrences data points. Root mean square error (RMSE) has been calculated based on the actual test data and predicted data using test data as input data after training the ANN. The observed RMSE value is 0.75. Figure 6 is showing the receiver operating characteristic (ROC) curve of the model which has produced the area under the curve (AUC) of 0.722. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold setting and therefore, it is used for model performance assessment. Higher the value of AUC, higher the prediction power of the model.

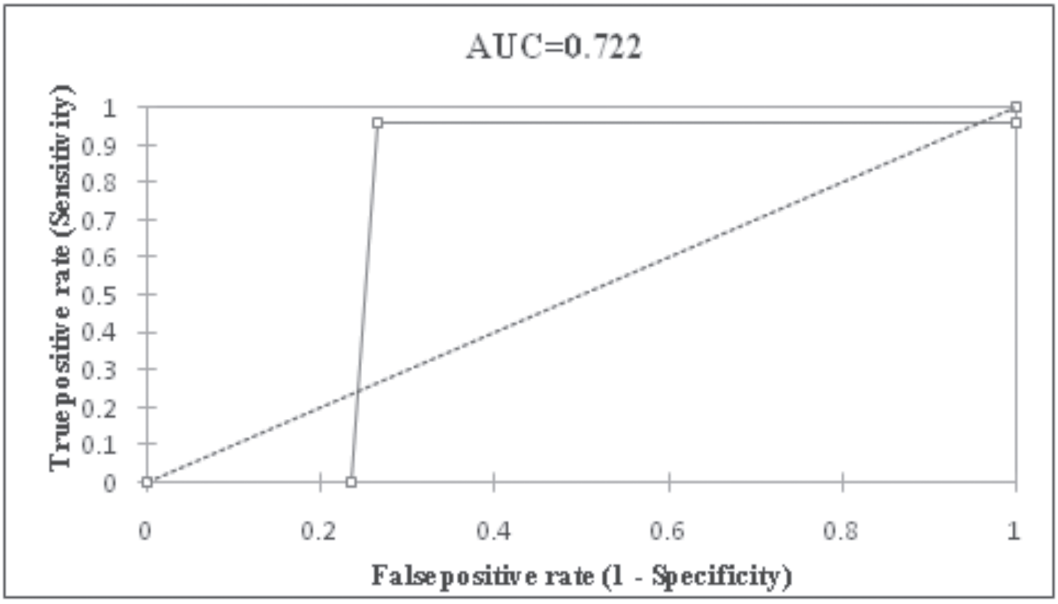

Figure 6: ROC for Validation of RBFN Model

Also, accuracy score of the model has been determined that is 0.65. In multi-label classification, the function returns the subset accuracy. The accuracy score is 1, that is the maximum value of accuracy score which can be achieved only when the set of predicted labels strictly match with the true set of labels and it is mathematically expressed in the equation 7 where ((y_i) is the predicted value of the i^{th} sample and y_i is the respective true value, then the fraction of true predictions over n_samples is calculated.

$$\text{accuracy}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{sample}}-1} 1(\hat{y_i} = y_i) \quad (7)$$

Figure 7(a) is showing risk map of DF in Delhi, and the spatial extent is ward-wise. Figure 7(b) is showing the wards which have been taken to prepare training dataset. The level value 0 is showing those wards which have not been taken in training dataset. Figure 7(c) is showing the

output of the model that is the predicted level of DF occurrences risk in each ward which has been taken to prepare test data in RBFN model. The wards with red colour are those wards which have very high risk of DF occurrences.The level value 0 is showing those wards which have not been taken in testing dataset. In the prediction map, the risk in most of the wards have been classified accurately, and very few wards have been misclassified when compared with risk map.

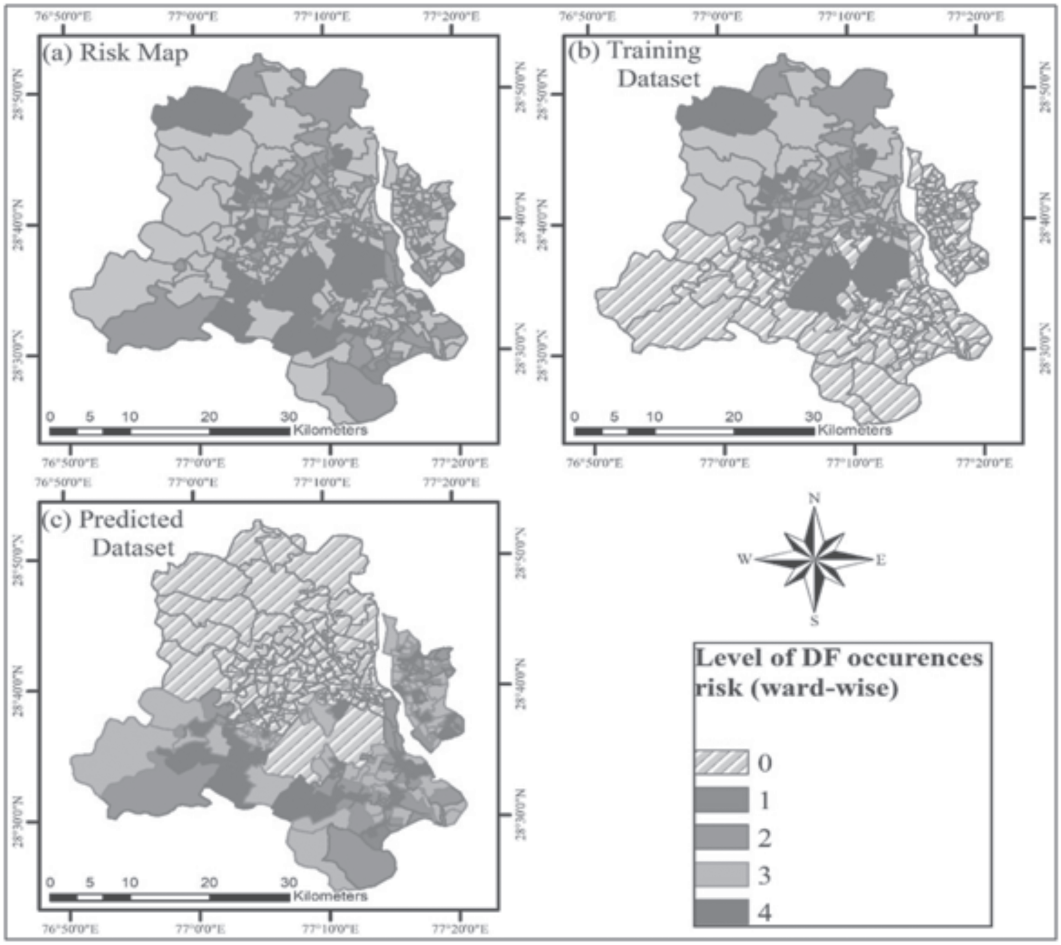

Figure 7: Dengue Fever Susceptibility Maps (Ward-wise)

In the present study, according to the results of logistic regression, it is evident that meteorological parameters are affecting the spatial as well as the temporal distribution of DF incidences.The impact of temperature has been associated with an increase in the DF incidences because of the increase in larval abundance which is in line with the studies performed (Pinto et al., 2011; Kumar et al., 2015). Also, by analysing the DF susceptibility maps, it can be observed

that ANN technique such as RBFN is effective in the generation of these maps with good accuracy. This study is useful in determining the potential areas of DF occurrences particularly in rural areas where early warning of DF would help to take immediate actions to control its spread and also fast medical facilities could be provided by identifying the locations of DF occurrences.

Conclusion

The association of DF incidences with meteorological parameters in the study area Delhi has been analysed and the study period is 2006-2013. In this study, strong association has been found between level of DF occurrences risk and meteorological parameters such as minimum temperature and absolute humidity. ANN technique such as RBFN has been used to produce DF susceptibility maps which could help identify potential areas of high risk of DF incidences. GIS has been used to produce susceptibility maps and to determine the areas geographically which are sensitive to DF. The information communicated by our analysis can be used by public health officials, epidemiologists and public health policymakers to plan the control and preventive measures for the occurrences of DF cases, especially for rural areas. Also, to monitor and measure the effectiveness of such strategies.

References

Hii, Y. L., Rocklöv, J., Ng, N., Tang, C. S., Pang, F. Y. and Sauerborn, R. (2009), 'Climate Variability and Increase in intensity and Magnitude of Dengue Incidence in Singapore', *Global Health Action,* 2(1), pp. 2036.

Kumar, S., Singh, M. and Chakraborty, A. (2015), 'Climatic Imbalance and their Effect on Prevalence of Dengue Fever in India', *International Journal of Current Microbiology and Applied Sciences*, 4(11), pp.185-191.

Lee, S. and Sambath, T. (2006), 'Landslide susceptibility Mapping in the Damrei Romel area, Cambodia Using Frequency Ratio and Logistic Regression Models', *Environmental Geology*, 50(6), pp. 847–855.

Looney, C. G. (2002), 'Radial Basis Functional Link Nets and Fuzzy Reasoning', Neurocomputing. *Elsevier*, 48(1–4), pp. 489–509.

Pinto, E., Coelho, M. and Oliver, L. (2011), 'The Influence of Climate Variables on Dengue in Singapore', *International Journal of Environmental Health Research*, 21(6), pp.415-426. Tsegaw, T., Gadisa, E., Seid, A., Abera, A., Teshome, A., Mulugeta, A., Herrero, M., Argaw, D., Jorge, A. and Aseffa, A. (2013) 'Identification of Environmental Parameters and Risk Mapping of Visceral Leishmaniasis in Ethiopia by Using Geographical Information Systems and a Statistical Approach', *Geospatial Health*, 7(2), pp. 299–308.

Ughelli, V., Lisnichuk, Y., Paciello, J., Pane, J. (2017), 'Prediction of Dengue Cases in Paraguay Using Artificial Neural Networks', The 3rd Int'l Conf on Health Informatics and Medical Systems.

Vikram, K., Nagpal, B. N., Pande, V., Srivastava, A., Gupta, S. K., Paul, R., Valecha, N. and Telle, O. (2015), 'Comparison of Ae . Aegypti Breeding in Localities of Different Socio-economic Groups of Delhi , India', *International Journal of Mosquito Research*, 2(2), pp. 83–88.

WHO (2013), 'Factsheet: Dengue and Severe Dengue'.

WHO (2014), 'Factsheet: Dengue and Severe Dengue'.

Xu, H., Fu, X., Lee, L., Ma, S., Goh, K. and Wong, J. (2014), 'Statistical Modeling Reveals the Effect of Absolute Humidity on Dengue in Singapore', PLOS Neglected Tropical Diseases,8(5), pp.e2805.

Yang, C. (2016) A Comparison of Four Methods of Diseases Mapping. Department of Physical Geography and Ecosystem Science, Lund University,Sweden.