Developing a Statistical Dengue Risk Prediction Model for the State of Delhi Based on Various Environmental Variables

Bhardwaj, A.,[1] Sam, L. C.,[1] Joshi, P. K.[1] and Sinha, V. S. P. [1,2]
[1]Department of Natural resources, TERI University, New Delhi, India, E-mail: pkjoshi27@hotmail.com
[2]Department of Remote Sensing, Banasthali University, Rajasthan, India

Abstract
This work investigates dengue affected localities of Delhi using static and dynamic environmental factors and their possible spatial relationships. The static variables include soil drainage, built-up area and vegetation. The dynamic variables represent seasonal precipitation and temperature data for past hundred years. Significance test (t-test) provided deterministic evidence of variable importance to model. Weighted sum and quantile classification helped to create a final risk map. The model indicated non-uniform distribution of risk across the state and showed elevated risk in urban built-up areas mainly alongside the river Yamuna. Three years (2007, 2008 and 2009) data for confirmed dengue cases for affected localities were obtained from Municipal Corporation of Delhi (MCD) for validation. 57.98% of the reported cases were observed under high risk category as modeled in this study. Modeling results indicate that environmental factors like precipitation, temperature, soil drainage, built-up area and vegetation govern mosquito breeding and are correlated with human dengue risk. The approach verified that dengue risk can be modeled at the state level and can be modified for risk predictions of other vector-borne diseases in varied ecological regions.

1. Introduction

The subject of disease outbreak prediction modeling has long been of help to community and public health planners (Comber et al., 2011). Previous research on outbreak prediction has been in two distinct and usually non-overlapping areas. One branch has considered the spatial dimensions related to geographical features (vegetation, water bodies, etc.), with data being manipulated and geographically analyzed using Geographical Information Systems (GIS) before subsequent statistical analyses (Cooke et al., 2006). Another body of research has examined outbreak prediction by considering the socioeconomic aspects with data collected using opinion or attitudes surveys (Bhandari et al., 2008). However, the emergence, re-emergence and distribution of vector-borne diseases are controlled both by geographic as well as socioeconomic factors such as structure of ecosystems (Clennon et al., 2010), climatic variability (Debien et al., 2010), human behavior (WHO, 2005), and ecology of vector and animal hosts of the infectious agent (Roels et al., 2011). Analyzing these complex landscape elements of interacting agents requires an interdisciplinary approach. Data from different sources and at different scales need to be linked, using geospatial analytic methods (Lambin et al., 2010). There is an emergent need to focus on correlating the health data with the spatial dimensions using geospatial tools to develop a valid and acceptable outbreak prediction model for vector borne diseases. Dengue is an arboviral infection and is of great concern nowadays (Solomon et al., 2000). World Health Organization (WHO) currently estimates 50 million dengue infections worldwide every year. The incidence of dengue has grown dramatically around the world in recent decades as some 2.5 billion people – two fifths of the world's population – are now at risk from dengue (WHO, 2009). The disease is now endemic in more than 100 countries in Africa, America, Eastern Mediterranean, Western Pacific and South-east Asia (most seriously affected) with an estimated 500,000 people requiring hospitalization each year (a very large proportion of whom are children) and about 2.5% of those affected die (WHO, 2009). India is one of the countries which had acute incidences of Dengue in recent past. New Delhi, the capital of India, has emerged as one of the hotspots of Dengue and Dengue Hemorrhagic Fever (WHO, 2007).

This study is taken up to estimate likelihood of dengue infection in the state of Delhi by analyzing spatial and climatic data to model habitat suitability for the vector, *Aedesaegypti*. It inhabits in a variety of environments and can be found in urban as well as rural settings. We viewed mosquito habitat suitability as a surrogate for estimating potential risk of dengue infection for humans and tested the usefulness of selected environmental variables in an analytical risk model. The output was in accordance with the environmental conditions promoting the growth of mosquitoes and the health data obtained from the health centers. The districts of East Delhi, North East Delhi, Central Delhi and North Delhi are amongst most severely affected. The areas are adjoining to river Yamuna.

2. Methodology

2.1 Study Area

Delhi occupies an area of approximately 1463sq km of which around 800 sq km (around 52%) is classified as urban (Rahman et al., 2011). The city is divided into nine districts (figure 1). Delhi remains typically hot in summer, humid in monsoon and cold during the winters. The average temperature of New Delhi ranges from 25°C to 46°C during summer and 2°C to 25°C during winter with an average annual rainfall of 570 mm (http://indiawaterportal.org). The cold waves from the Himalayan region along with winter rains due to western disturbances make the winters very chilly. In summers, the heat wave is enormous and adequate precaution has to be taken before going out in the afternoons. Such big and sudden climatic variations make it a hotspot for vector borne diseases. Three agencies, namely Municipal Corporation of Delhi (MCD), New Delhi Municipal Committee (NDMC) and Ministry of Defense are responsible for dengue control activities in the state.

2.2 Data Used

Three years (2007, 2008 and 2009) locality-wise monthly disease data were obtained from the MCD. The geographic coordinates of these localities were obtained using Google earth. The year was divided into four seasons (rainy, summer, winter and spring) using temperature and rainfall observations of previous hundred years (http://indiawaterportal.org). The built-up area (figure 2a) and vegetation maps (figure 2b) were prepared using October 2010 Landsat ETM satellite imagery and high resolution Google earth imagery. Soil map (1:125000) was digitized from the database prepared by National Bureau of Soil Survey & Land Use Planning (NBSS & LUP) and the drainage network from Survey of India (SOI) toposheet (figure 2c). Seasonal dengue occurrence maps were prepared from Inverse Distance Weighted (IDW) interpolation of the point data using number of confirmed cases (figure 3a-d).

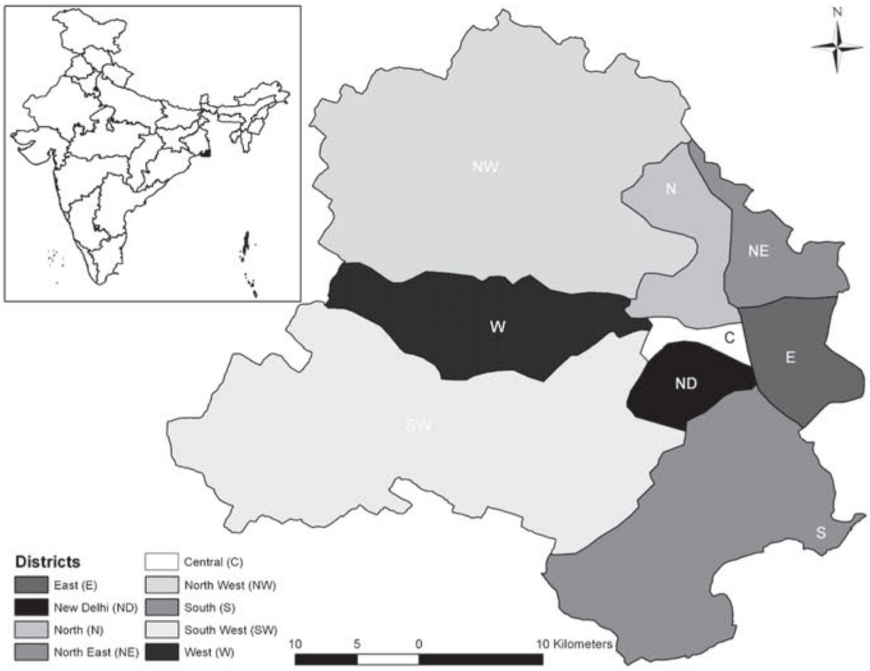

Figure 1: Location of the Study Area

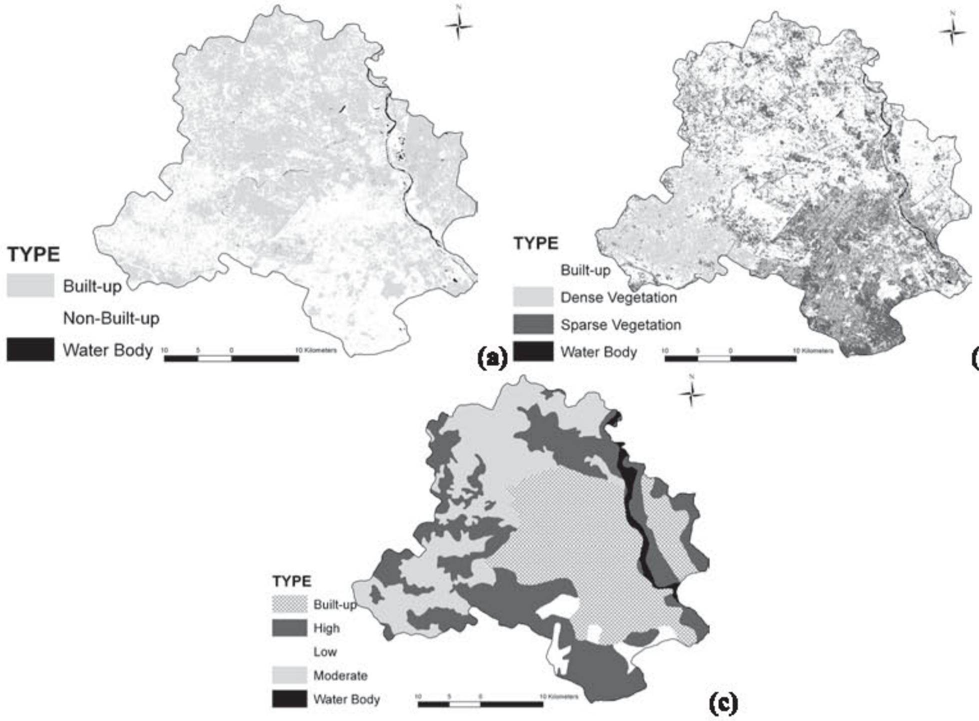

Figure 2: Distribution of (a) Habitation, (b) Vegetation and (c) Soil drainage

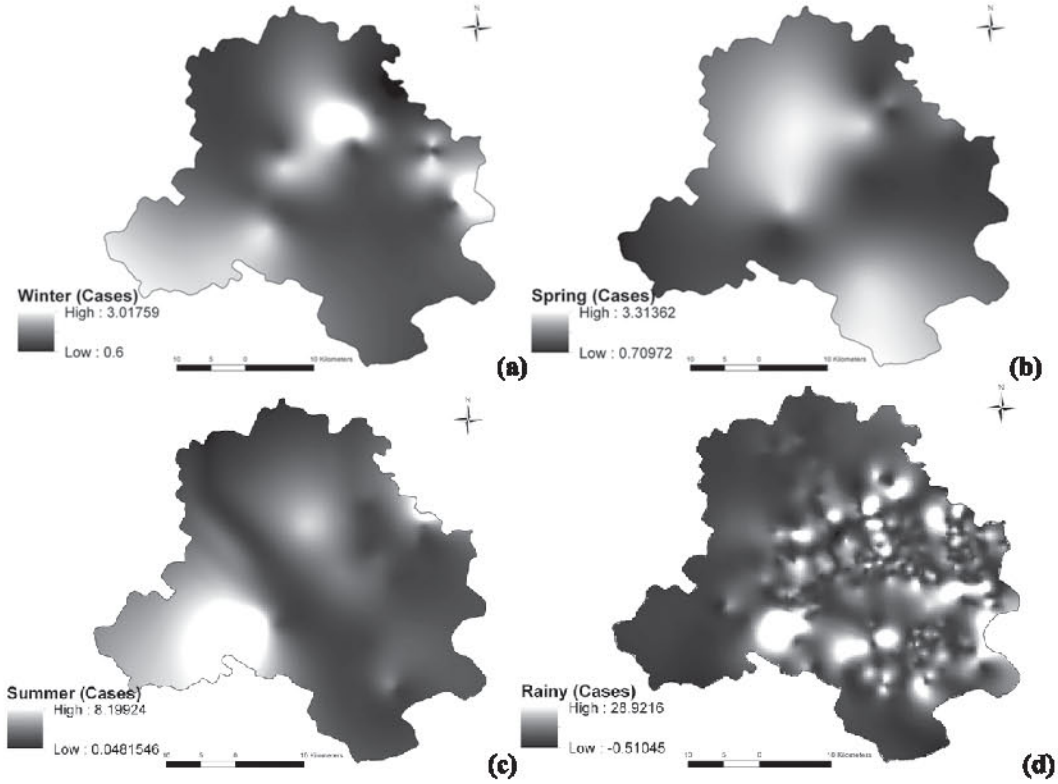

Figure 3: Distribution of Dengue cases in different seasons (a) Winter, (b) Spring, (c) Summer and (d) Rainy

2.3 Spatial Modeling

Broadly the methodology followed was preparation of maps of static and dynamic variables, integration of layers using statistical model in GIS environment, prediction of outbreak risk as a map and validation of prediction. The selection of environmental variables used in the study was based on evaluation of specific mosquito habitat conditions (Kolivras, 2006) and conditions promoting mosquito breeding (Winch et al., 2002). The static variables were assumed to change slowly or not at all over the time. The static variables chosen were soil drainage, built-up area and vegetation. The dynamic variables selected were climatic and meteorological conditions throughout the year. The potential multi-colinearity among the variables was disregarded and it was assumed that all are important to the dengue risk estimation. Ranks to the static layers were assigned by performing t-test. The *t*-test compares one variable (in this case, occurrence of dengue) between two groups (here built-up and non-built-up, vegetated and non-vegetated, and moderately and excessively drained soil classes). Students' t-test was used to assess whether the means of two groups are statistically different from each other. More the t value, more significant is the difference between the means. Given below is the formula to calculate t-value:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Formula 1)

Where,

$\bar{x}_1$ = Mean of sample 1

$\bar{x}_2$ = Mean of sample 2

n_1 = Number of subjects in sample 1

n_2 = Number of subjects in sample 2

s_1^2 = Variance of sample 1

s_2^2 = Variance of sample 2

A flow diagram (figure 4) of the methodology followed is given below:

Based on the t-test, it was observed that the most significant factor for the occurrence of dengue was soil drainage (highest t value) followed by built-up and vegetation. So the respective ranks given to these layers were 1 for soil drainage, 2 for built-up and 3 for vegetation. The weights were assigned using following formula (Cooke et al. 2006):

$$w_j = \frac{n - r_j + 1}{\Sigma(n - r_k + 1)}$$

(Formula 2)

Where, w_j is the normalized weight for the j^{th} variable:

n is the number of variables under consideration

r_j is the rank position of the variable

Using formula 2, the weights for different static layers of significance were computed. The calculated weights were 0.5 for soil drainage, 0.34 for built-up and 0.16 for vegetation. The details of different classes and respective class weights are given in table 1. Ranks and weights to the dynamic seasonal layers were assigned based on the percentage of confirmed cases in that season. For example around 80% cases were recorded in rainy season. So it was assigned a weight of 0.8. Likewise summer, winter and spring were assigned weights of 0.1, 0.06 and 0.04 respectively. Assigning the ranks to these layers did not require t-test as the difference between number of confirmed cases recorded in rainy season and those recorded in other three seasons was huge, thus proving the importance of precipitation and temperature characteristics of rainy season over other dynamic and static layers favoring dengue occurrences. So in order to incorporate this prevalence in the model, rainy season was assigned the highest weight.

Table 1: Distribution of different classes and respective class weight

	Habitation				Vegetation				Soil Drainage			
Grid Code	Class	Area (sq km)	Area (%)	Class weight	Class	Area (sq km)	Area (%)	Class weight	Class	Area (sq km)	Area (%)	Class weight
1	Water body	9.71	0.65	1	Water body	9.71	0.65	1	Water body	25.07	1.83	1
2	Built-up	770.16	51.87	5	Built-up	770.16	51.87	5	Built-up	437.42	32.00	5
3	Non built-up	704.83	47.47	3	Sparse vegetation	450.61	30.35	4	Moderate	368.25	26.94	4
4					Dense vegetation	250.75	16.89	3	Low	45.00	3.29	2
5									High	491.20	35.93	3

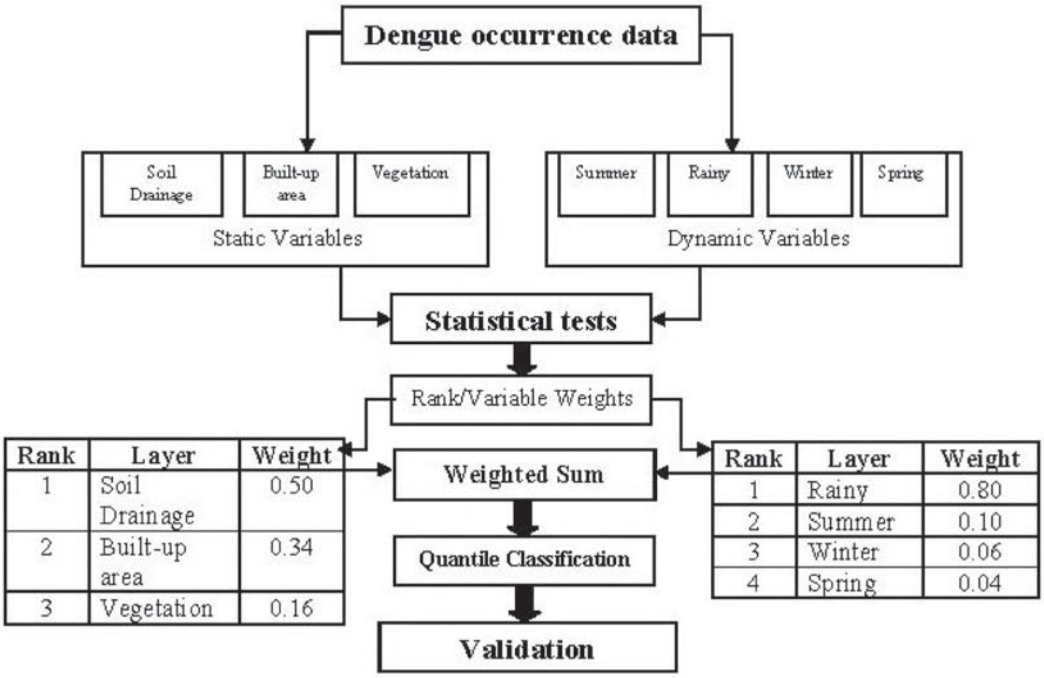

Figure 4: Flow chart showing methodology followed

2.4 Weighted Sum and Quantile Classification

ArcGIS provides weighted sum tool to weigh and combine multiple inputs to create an integrated analysis model. It is similar to weighted overlay tool in which multiple raster inputs representing multiple variables can be easily combined incorporating weights or relative importance. One major difference between weighted overlay tool and weighted sum tool is that weighted sum tool allows for floating point values whereas weighted overlay tool only accepts integer values (http://webhelp.esri.com). Generally, the values of continuous rasters such as slope or Euclidean distance outputs are grouped into ranges, where each range is assigned a single value to represent a class such as low, medium or high importance. The reclassify tool allows reclassifying the values for such rasters. The weighted overlay tool is used most commonly for suitability modeling and should be used to ensure that the correct methodologies are followed. The weighted sum tool is useful when floating-point output or decimal weights are required. The following model equation was used in the weighted sum tool:

Outbreak risk = Static model + Dynamic model

Where, Static model = Soil drainage × 0.5 + Built-up × 0.34 + Vegetation × 0.16

Dynamic model = Rainy × 0.8 + Summer × 0.1 + Winter × 0.06 + Spring × 0.04

The output was further refined into five risk categories by quantile classification. Quantile is one of the class of values of a variate which divides the members of sample into equal-sized subgroups of adjacent values or a probability distribution into distributions of equal probability (http://support.esri.com).

3. Results

The outbreak prediction map was obtained after following the above mentioned model. In Delhi, the dengue risk appears to be associated mainly with rainy season and subsequent water logging because of poor soil drainage. The dengue susceptibility map is shown in figure 5. The areas recorded for different risk classes are shown in table 2. Dense built-up was also found to be significant because of higher population at risk. Looking at the risk map keeping in mind both dense built-up and water logging factors, we observed that the zones along the Yamuna river either fall under high risk or elevated risk. The districts of East Delhi, North East Delhi, Central Delhi and North Delhi are most densely inhabited and are worst hit. This region records maximum number of cases each year.

This output was compared with the MCD health data for consecutive three years (2007, 2008 and 2009). Very high degree of correlation was observed. About 58% of total cases fall in high risk zone of the output followed by 33% in elevated risk zone. The areas under low and moderate risks accounted only for 2% of total cases thus supporting the model hypothesis strongly. These low risk areas were observed mainly along the outskirts of the city.

Table 2: Distribution of different risk classes

Class	Area (%)	Recorded cases (%)
High risk (5)	19.73	57.98
Elevated risk (4)	21.43	33.02
Moderate risk (3)	20.48	7.00
Low risk (2)	19.66	1.20
Minimal risk (1)	18.70	0.80

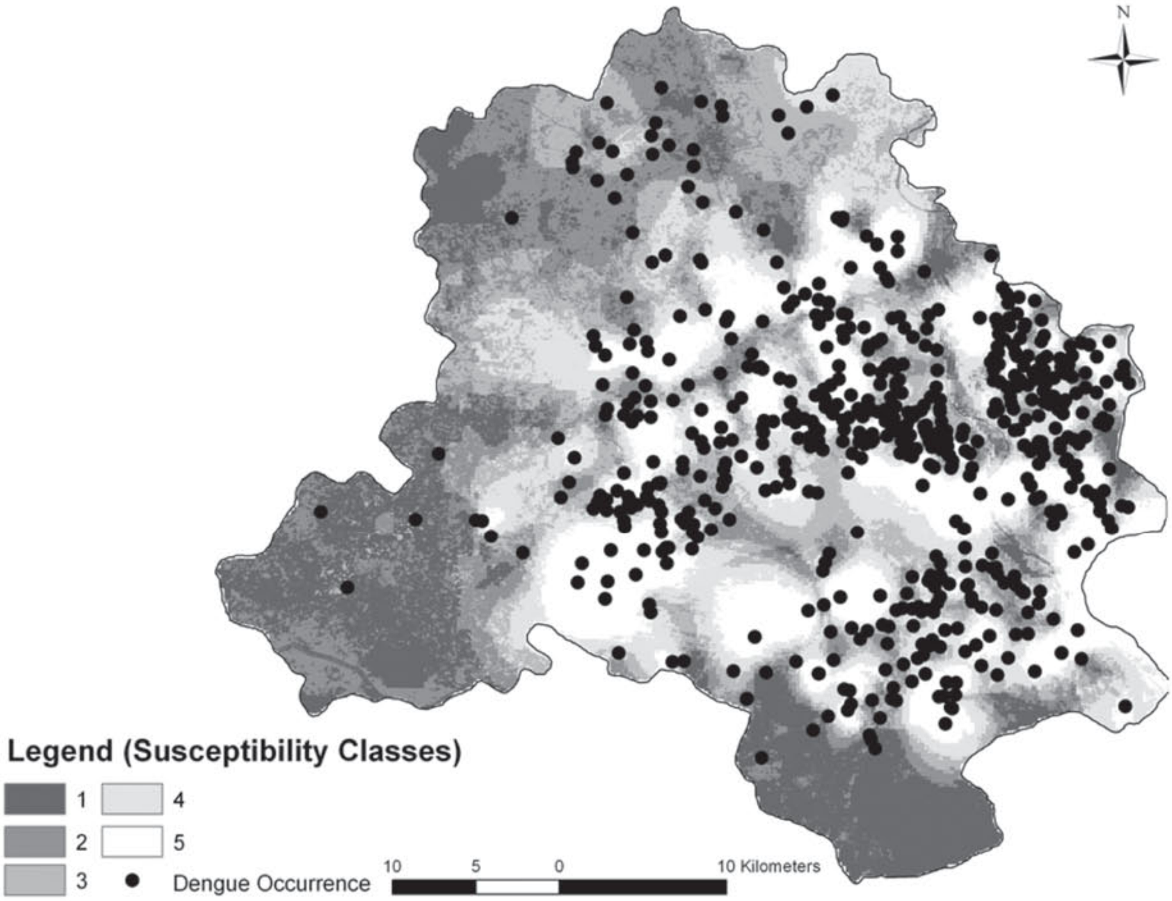

Figure 5: Result and Validation

4. Discussion and Conclusions

In this study, spatial estimation of dengue risk in Delhi was carried out by analyzing dengue and environmental data. The model was developed assuming that mosquito habitat suitability factors can be used to estimate the risk of dengue. The t-test for each pair of extreme variables was the basis for ranking the static layer importance. Several environmental factors were considered and according to our analysis, in Delhi, dengue risk is correlated to high built-up (0.34) and moderate drainage of soil (0.5) as well as occurrence of rains and water logging (0.8). The methodology developed is exclusively based on health data and can be easily modified for various vector-borne diseases in varied ecological regions. Dengue risk map was validated with human case data and clearly shows areas environmentally prone to sustaining the virus. The vegetation layer was included in the model based on literature survey as it is supposed to be a good habitat for mosquito breeding but it was found to be of least importance as majority of population resides in dense built-up areas of the city and away from vegetated areas. In a city like Delhi, most of the population resides within densely built-up area. This was the reason behind more number of cases from built-up areas. Also the areas along the river Yamuna fall under high and elevated risk categories. This was further confirmed with the MCD dengue data. The outcome could be refined if the health data available would have covered all types of medical facilities including the MCD centers. Also the soil drainage map used was generalized and older (1999) for such a study. These were the main reasons behind a large area difference of built-up in soil drainage map and built-up in other static layers of vegetation and habitation as the Landsat images used to prepare these layers were of 2010. Over the span of these eleven years, Delhi has experienced unforeseen constructional activities leading to advent of two of the largest sub-cities of Asia, Dwarka and Rohini. Heuristic methods of variable weighting often employed in GIS analyses can introduce personal bias in the modeling process (Cooke et al. 2006). In our study, statistical tests of environmental variable significance provided deterministic evidence of each variable importance (weight) for predicting risk using GIS. This approach diminishes the possibility of introduction of analyst bias in models (Cooke et al., 2006). This information can help to develop mosquito control strategies and aid regulatory agencies to prioritize their prevention efforts.

References

Bhandari, K. P., Raju, P. L. N. and Sokhi, B. S., 2008, Application of GIS Modeling for Dengue Fever Prone Area Based on Socio-Cultural and Environmental Factors – A Case Study of Delhi City Zone. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B8. Beijing:* 165-171.

Clennon, J. A., Kamanga, A., Musapa, M., Shiff, C. and Glass, G. E., 2010, Identifying Malaria Vector Breeding Habitats with Remote Sensing Data and Terrain-Based Landscape Indices in Zambia. *International Journal of Health Geographics*, 9(58),

Comber, A. J., Brunsdon, C. and Radburn, R., 2011, A Spatial Analysis of Variations in Health Access: Linking Geography, Socio-Economic Status and Access Perceptions. *International Journal of Health Geographics*, 10(44),

Cooke, W. H., Grala, K. and Wallis, R. C., 2006, Avian GIS Models Signal Human Risk for West Nile Virus in Mississippi. *International Journal of Health Geographics*, 5(36),

Debien, A., Neerinckx, S., Kimaro, D. and Gulinck, H., 2010, Influence of Satellite-Derived Rainfall Patterns on Plague Occurrence in Northeast Tanzania. *International Journal of Health Geographics*, 9(60),

Dengue and dengue haemorrhagic fever Fact Sheet N°117, March 2009. http://www.who.int/mediacentre/factsheets/fs117/en/ (Accessed 5 August 2010). http://indiawaterportal.org (Accessed 12 September 2010)

Kolivras, K. N., 2006, Mosquito Habitat and Dengue Risk Potential in Hawaii: A Conceptual Framework and GIS Application. *The Professional Geographer, 58(2)*, 139–154

Lambin, E. F., Tran, A., Vanwambeke, S. O., Linard, C. and Soti, V., 2010, Pathogenic Landscapes: Interactions between Land, People, Disease Vectors, and their Animal Hosts. *International Journal of Health Geographics*, 9(54),

Rahman, A., Kumar, Y. and FazalS, B. S., 2011, Urbanization and Quality of Urban Environment using Remote Sensing and GIS Techniques in East Delhi-India. *Journal of Geographic Information System*, 3, 61-83.

Roels, S. B., Peterson, A. T. and Xiao, X., 2011, Comparative Analysis of Remotely-Sensed Data Products via Ecological Niche Modeling of Avian Influenza Case Occurrences in Middle Eastern Poultry. *International Journal of Health Geographics*, 10(21),

Solomon, T., Dung, N. M., Vaughn, D. W., Kneen, R., Thao, L. T. T,, Raengsakulrach, B., Loan, H. T., Day, N. P. J., Farrar, J., Myint, K. S. A., Warrell, M. J., James, W. S., Nisalak, A. and White, N. J., 2000, Neurological Manifestations of Dengue Infection. *The Lancet. Vol: 355.* 1053–59.

Winch, P. J,, Leontsini, E., Perez, J. G. R., Perez, M. R., Clark, G. G. and Gubler, D. J., 2002, Community-Based Dengue Prevention Programs in Puerto Rico: Impact on Knowledge, Behavior, and Residential Mosquito Infestation. *American Journal of Tropical Medicine and Hygiene,* 67(4), 363–370.

World Health Organisation, Regional Committee for the Eastern Mediterranean, August 2009, *Vector-Borne Diseases: Addressing a Re-Emerging Public Health Problem.* Fifty-second Session Agenda item 5 (a),

World Health Organisation, 2007, Regional Office for South-East Asia, September 2007, *Communicable Disease Newsletter*, 4(3). New Delhi.