Geo-statistical Dengue Risk Model Case Study of Lahore Dengue Outbreaks 2011

BILAL TARIQ
Department of Remote Sensing & Geo-information Science
Institute of Space Technology (IST)
Karachi Campus, Pakistan
Email: bilaltariq8@hotmail.com

ARJUMAND Z. ZAIDI
Assistant Professor, Department of Remote Sensing & Geo-information Science
Institute of Space Technology (IST),
Karachi Campus, Pakistan
Email: arjumand.zaidi@ist.edu.pk , arjzaidi@gmail.com

Abstract **-- Repetitive dengue fever outbreaks in Pakistan have brought major concerns to the government authorities to control this mosquito borne disease. Although dengue cases are reported in many cities of Pakistan but Lahore has been the most affected city during last few years. In 2011, dengue outbreaks in Lahore are considered as the worst epidemic in the national history. There is an urgent need to manage this disease effectively. In order to keep the virus under control, it is also required to explore the possible causes and factors that support dengue virus to grow. In this study climatic and environmental factors that may presumably promote the growth of virus are selected and their spatial and temporal variations are correlated with dengue cases. The objective of this paper is to develop a geospatial dengue risk model to identify the risk prone areas by linking these factors with dengue outbreaks using satellite data and Geographical Information System (GIS) techniques. Satellite images of SPOT-5, Landsat-TM and Google Earth are used in this study to derive environmental and landuse parameters. The model parameters used for this study *are Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), built-up area, population, population density* and *precipitation*. Ordinary Least Square (OLS) and Geographical Weighted Regression (GWR) analyses are performed to develop regression models between dengue cases and other study parameters. Based on study results, it is concluded that study parameters are not suitable for OLS global model since no statistically strong model can be found using OLS. GWR analysis is a form of linear regression that can model spatially varying relationships between variables. The GWR model shows that using *population density* and *built-up area* as explanatory variables, the model can explain 77% of the variance in dengue incidences.**

Keywords: Dengue Fever; Risk Mapping; Geo-statistics, Geographical weighted regression, Ordinary Least Square Regression

I. Introduction

Mosquito-borne diseases are becoming a major threat to public health all over the word especially in tropical and sub-tropical regions. Dengue fever is amongst the major mosquito-borne diseases besides malaria. At present, dengue has become the most vital mosquito-borne viral infection in the world [1]. Severe forms of dengue fever, *Dengue Haemorrhagic Fever (DHF)* and *Dengue Shock Syndrome (DSS)* are considered life threatening and fatal diseases.

In Pakistan, the first dengue case was registered some twenty years back in 1994 in Karachi. Since then many outbreaks have been reported all over the country during monsoon seasons. Lahore experienced the worst episode of dengue outbreak in 2011 that started in March and prevailed till December. Majority of the cases were reported during the hot and humid months of August and October. In 2011, more than seventeen thousand (17,000) cases in Lahore were registered in local hospitals that were acquired for this study from Disaster Management Authority (PDMA) database.

As mentioned earlier dengue outbreaks are predominant in many hot and humid areas and therefore it is anticipated that climatic conditions may have some influence on the distribution pattern of dengue occurrences. Elevated temperature and humidity are considered suitable for mosquito breeding that may cause spread of dengue virus. Urban areas with high population density usually face rapid transmission of communal diseases and therefore besides climatic variables, landuse/cover may also be considered as an influencing parameter for dengue outbreaks.

To study dengue distribution pattern, Geographical Information System (GIS) can be utilized as an efficient tool to manage and analyze spatial and temporal data and establish their relationships among study parameters. GIS is being used in many parts of the world for monitoring and mitigating epidemics and to do rapid mapping of risk prone areas [2, 3, 4]. GIS techniques combined with Satellite Remote Sensing (RS) may also provide a wealth of data derived from satellite images and analysis of that data in a meaningful way.

II. Objective

The main objective of this project is;

To develop geostatistical dengue risk model for identification of risk prone areas by linking environmental, demographic and landuse/cover parameters with dengue cases using satellite data and Geographical Information System (GIS) techniques. With the help of this model the high risk areas of Lahore can be identified. Model dependent variable is the number of dengue cases, whereas, independent or explanatory variables selected for this study are; *vegetation cover (NDVI), rainfall,* *land surface temperature (LST), built-up area, population* and *population density*. These factors are statistically tested for their contribution to dengue outbreaks by relating them with dengue cases.

A. Study Area

The most severe dengue outbreak in the history of the country occurred in 2011 in Lahore during the months of March and December (The News, 16 Aug 2011). Therefore, Lahore, the capital city of Punjab, is selected for this study to determine the risk prone areas of dengue fever outbreaks in the city. Lahore is considered as the second largest city of Pakistan with an estimated population of 8,592,000 (2010 estimates). The geographical location of the city is 31°32′59″N and 74°20′37″E and it is situated at an altitude of 217 meters above mean sea level. At its north flows the famous river Ravi and its eastern side is country’s border with India (Fig. 1).

Lahore is divided into nine administrative towns and one cantonment under 2001 revision of Pakistan's administrative structure as shown in Fig. 2. Lahore nine towns are Ravi Town, Shalimar Town, Wagah Town, Aziz Bhatti Town, Data Gunj Bakhsh Town, Gulberg Town, Samanabad Town, Iqbal Town, and Nishtar Town. These towns are further split into Union Councils (UC). A Union Council (UC) is the lowest administrative unit.

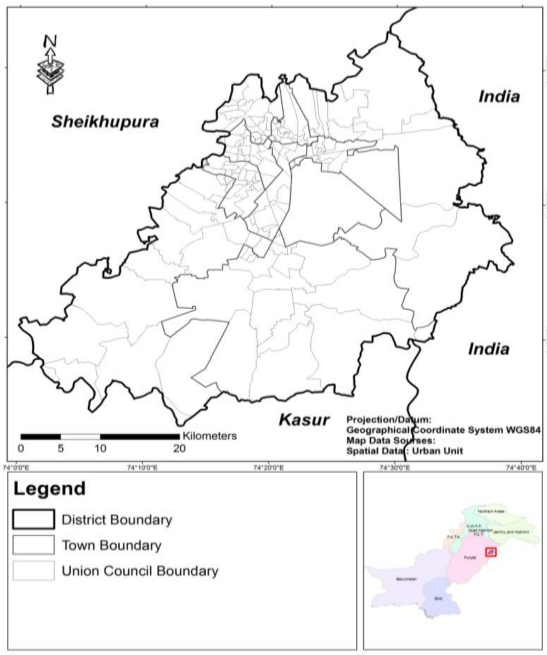


Figure 1. Location of Lahore City

III. MATERIALS AND METHODS

The methodology for developing geo-statistical dengue risk model is shown in Fig. 3. The study phases involved in the methodological workflow are; ***data collection*** (climatic, epidemic, satellite imageries and other relevant data), ***Geospatial analysis and Geo-database development***, and ***formulation of Geo-statistical model***.

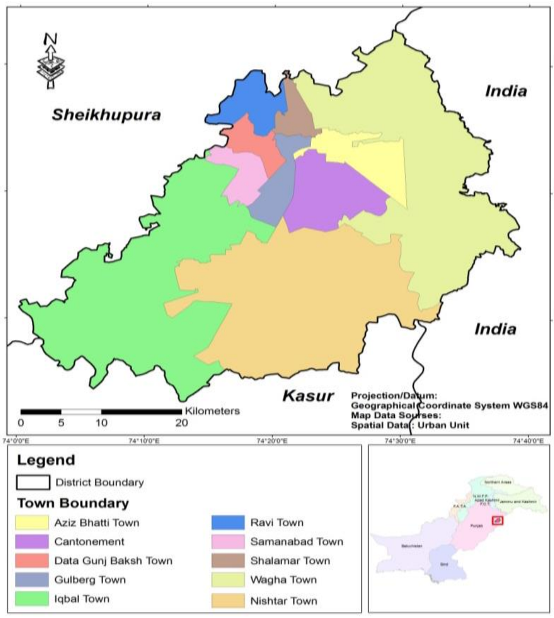


Figure 2. Nine Towns of Lahore

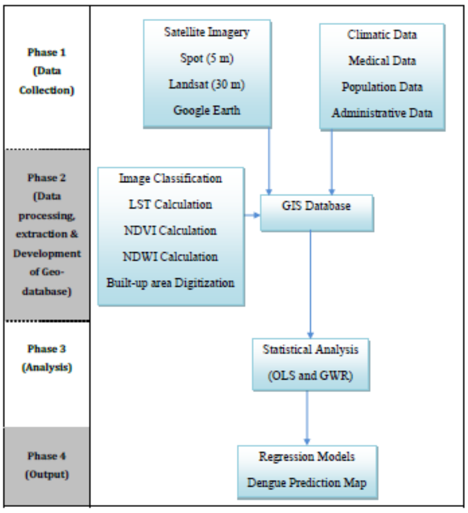


Figure 3. Methodological Concept

A. Data Collection

Satellite data from Spot-5, Landsat TM and Google Earth are used to derive study parameters including *Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Land Surface Temperature (LST)* and *built-up area*. *NDWI* and *NDVI* are obtained from Spot-5 satellite data; *LST* from Landsat TM satellite data and *built-up area* is digitized from Google Earth archive image. Precipitation and population data are acquired, respectively, from *Pakistan Meteorological Department* (PMD) and *Pre-investment Study* [5].

GIS tools are used for mapping and analysis of these parameters along with epidemic data. *Ministry of Health (MoH)* and *Punjab Disaster Management Authority (PDMA)* have compiled database of dengue cases registered at local hospitals. The same database, after data cleaning, is used in this study.

B. Geo-spatial Analysis and Geo-database Development

GIS software ArcGIS is used to digitize all collected data in a geo-database that is a common data storage and management structure for ArcGIS. A geo-database combines attribute data with spatial data for data management, processing and analysis. Thematic layers of study parameters are created and combined using different geo-processing and zonal statistics tools of ArcGIS software. Data conversion between raster and vector data models is also done wherever required.

C. Formulation of Geo-statistical Model

Statistical tools of ArcGIS including Ordinary Least Square (OLS) and Geographical Weighted Regression (GWR) are used in this study to develop geo-statistical model for dengue risk analysis. *NDVI, NDWI, LST, population density* and *built-up area* are used as independent variables that were regressed against the dependent variable - *number of dengue cases*. Linear regression analysis is also done for select UCs to model the impact of *precipitation* on dengue outbreak and using temporal data of rainfall and dengue cases during study period. Other independent variables are used to develop spatial model to identify risk areas. First, all variables are used at a time in the model and discarded one by one that show weak relationship with dengue cases. The variables that result in a statistically significant model are included in the risk model.

IV. RESULT AND DISCUSSION

Correlation and regression analysis are the major statistical tools used to find out the statistical significance in the relationship between dependent variable and explanatory variables. In this study, temporal weather data is used to find out the impact of temporal changes in weather parameters on dengue occurrences. The spatial variations of all other parameters are examined in a linear regression analysis to establish their relationships with spatial variation in dengue cases at union council level.

Influence of study parameters on dengue outbreaks is analyzed to select the most affecting ones for model development. The environmental and land cover factors are derived from satellite data and mapped using GIS techniques. SPOT 5 image of August 2011 is used for Land Cover Land Use (LCLU) classification. Although LCLU classification does not directly used in the analysis but it is helpful in acquiring the knowledge about land cover and land use features of the study area (Figure 4). The parameters that are used in this study to develop risk model are discussed in the following sections.

Land Surface Temperature

The *Land Surface Temperature (LST)* map is generated from thermal band of Landsat-5 (TM) satellite data that represents the distribution pattern of surface temperature in the study area. Post-incidence LST map for the month of August 2011 is used in this study when outbreak was at its peak. Zonal statistics of *LST* is derived using administrative boundary map where zones are defined by UC boundaries.

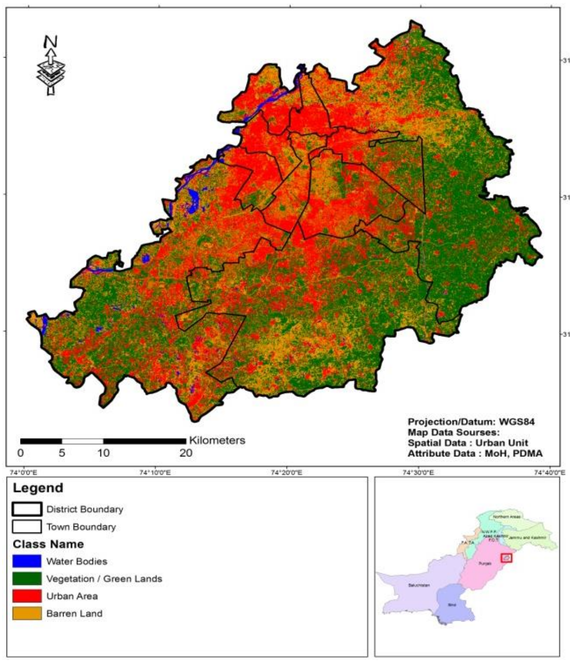


Figure 4. Land Cover and Land Use (LCLU) Map of District Lahore

Normalized Difference Vegetative Index and Normalized Difference Water Index

Normalized Difference Vegetative Index (NDVI) is a measure of vegetation cover on a land surface, whereas, *Normalized Difference Water Index (NDWI)* indicates the presence of water at the surface of land. SPOT-5 image is used to develop *NDVI* and *NDWI* maps. September 2011 satellite images are acquired for this purpose. Zonal statistics are calculated for *NDVI* and *NDWI* using UC shapefile that is used to define zonal boundaries.

BUILT-UP AREA

Built-up area is digitized from Google Earth archive image of august 2011. The digitized layer is converted to ArcGIS readable shapefile format. *Built-up area* within each UC is calculated by intersecting *built-up area* layer with UC layer and dissolving the resultant layer based on UC names. ArcGIS geo-processing tools '*intersect'* and '*dissolve*' are used for this purpose.

POPULATION AND POPULATION DENSITY

Population data were available at town level and it was required to further break it up at UC level since all analysis is done at this level. It is assumed that *built-up area* is an indicator of population and fraction of the total *built-up area* in each UC can be multiplied with town population to get UC level population. *Population density* is further calculated using UC level population divided by total UC area in square kilometer.

A. Statistical Analysis between Weather Parameters and Dengue Outbreaks

The weather parameters selected for this study are *precipitation*, *atmospheric temperature* and *relative humidity* in Lahore during study period. Stagnant waters are ideal sites for dengue breeding. In Lahore, after rainfall, water stands in areas that are relatively at lower elevation from the road network like open spaces, parks and green belts, graveyards and vacant plots. Female dengue mosquito lays eggs on clean stagnant water and the time period required from egg to become adult mosquito is 2-3 weeks depending upon the environmental conditions. In monsoon season, after first two or three rainfalls, the water absorbing capacity of saturated soil declines. Also the evaporation of surface water reduces due to high humidity and cloudy sky. Low infiltration along with a slow evaporation rate during the monsoon season results in standing water that lasts for several days. Due to these reasons, usually the dengue fever outbreaks occur during rainy seasons (refer Fig. 5 and Fig 6).

Since dengue cases usually occur after rainfall and the life span of dengue is 2-3 weeks, a date wise comparison of the two variables cannot be possible. To find out the relationship between these two factors, dengue cases are aligned with rainfall data by shifting them 15 to 18 days back from their recorded dates of incidence. Following time intervals are selected for shifting of dengue cases according to the amount of rainfall. Please note that if rainfall occurs for two or three consecutive days, then data is shifted in accordance with the total sum of the rainfall amount during these consecutive days.

- If rainfall is less than 1 mm, data are shifted 0-1 days from the registered dates.
- If rainfall is between 1.1-10 mm, data are shifted 2-4 days from the registered dates.
- If rainfall is between 10.1-40 mm, data are shifted 5-9 days from the registered dates.
- If rainfall is equal to 40.1 mm or greater, data are shifted 10-18 days from the registered dates.

Three towns (Data Gunj Baksh Town, Samanabad Town, Gulberg Town) are selected as sample cases to find out the temporal relationship between dengue incidences and weather parameters in select union councils of these towns. Firstly, scatter plot matrices are developed using ArcGIS graphical tool and the parameters that show a linear relationship with dengue incidences are analyzed further to develop linear regression model. In all cases only rainfall amount (*precipitation*) and dengue cases are selected for regression analyses. *Atmospheric temperature* and *relative humidity* data apparently cannot establish any significant relationship with dengue cases. Results of linear regression show statistically significant relationship between *precipitation* and dengue outbreaks in all sample union councils with R^2 ranging from 0.66 (minimum) to 0.8 (maximum) as shown in Figs. 7, 8, 9, 10 and 11.

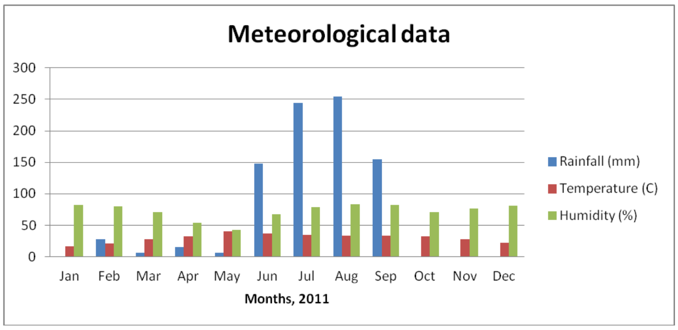


Figure 5. Meteorological data of District Lahore in 2011

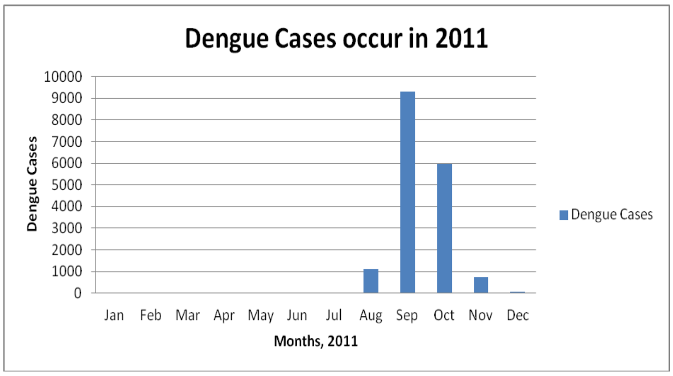


Figure 6. Dengue Incidences in District Lahore during 2011

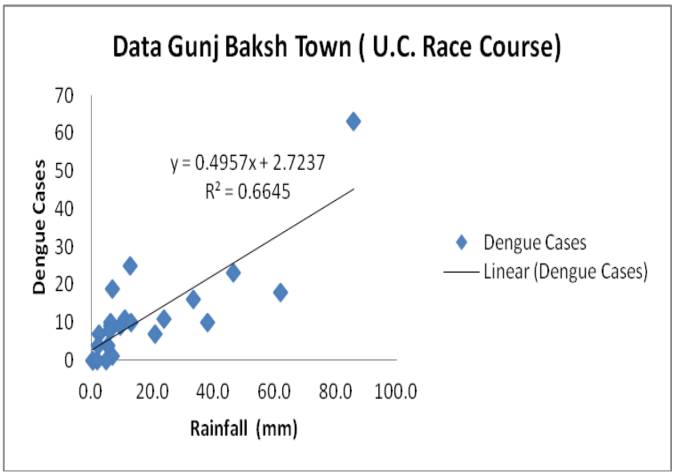


Figure 7. Scatter Plot and Regression Line between Rainfall and Dengue cases in U.C. Race Course, Data Gunj Baksh Town

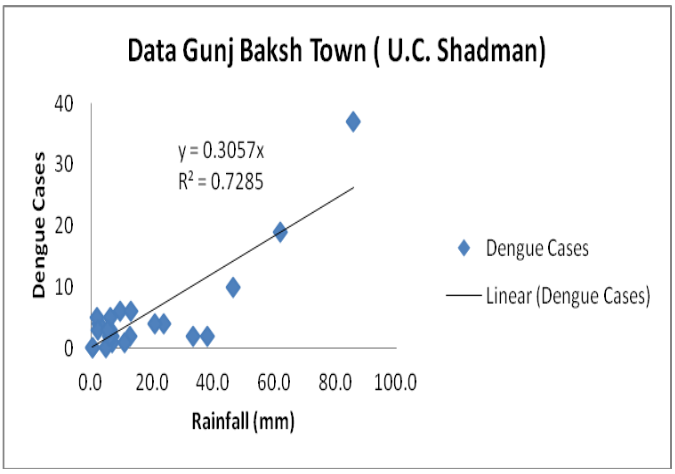


Figure 8. Scatter Plot and Regression Line between Rainfall and Dengue cases in U.C. Shadman, Data Gunj Baksh Town

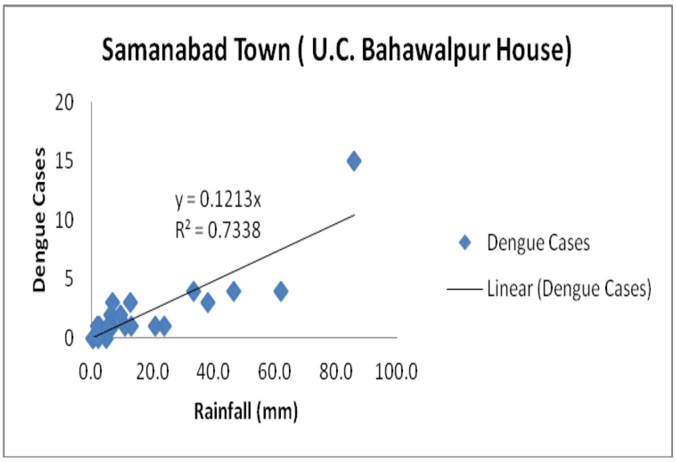


Figure 9. Scatter Plot and Regression Line between Rainfall and Dengue cases in U.C. Bahawalpur House, Samanabad Town

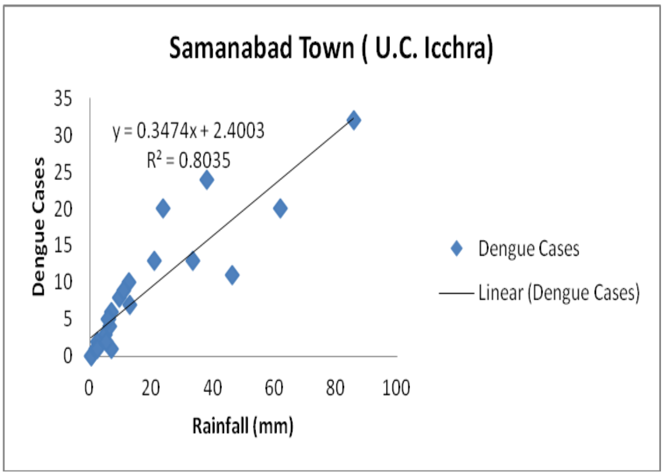


Figure 10. Scatter Plot and Regression Line between Rainfall and Dengue cases in U.C. Icchra, Samanabad Town

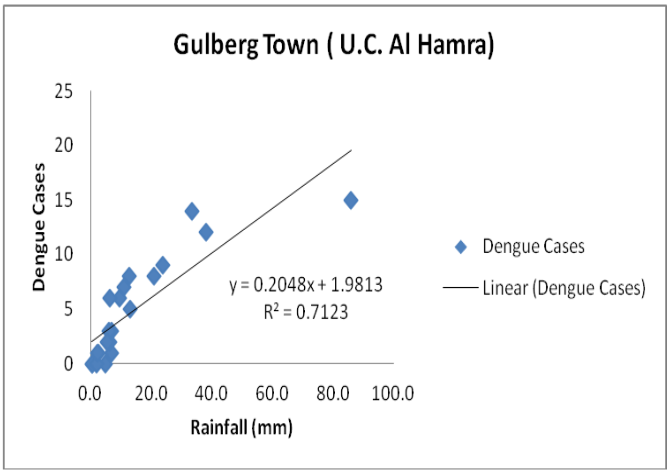


Figure 11. Scatter Plot and Regression Line between Rainfall and Dengue cases in U.C. Al Hamra, Gulberg Town

B. *Geo-statistical Risk Model*

In order to test the research hypothesis and to find out the relationship between dengue cases and study parameters and the direction of relations, the following regression analyses are conducted on the data.

i. *Ordinary Least Squares (OLS) Regression*
ii. *Geographically Weighted Regression (GWR)*

Ordinary Least Squares (OLS) is a global regression method. Geographically Weighted Regression (GWR) is a local spatial regression method that allows the relationships to model across the study area.

i. *OLS Regression between Dengue Cases and All Study Parameters*

Average values of *NDVI, NDWI, LST* and maximum *LST* are calculated at UC level using zonal statistics tool of ArcGIS and used in the statistical analysis. Other parameters used in this analysis are 2008 *population* (in thousands), *population density* and *built-up area*. Scatter plot matrix is developed in ArcGIS to get an idea if any ordinary linear relationship exists between dengue cases and other parameters or not? Although many parameters do not have apparent linear relationship with dengue cases but a trial OLS regression analysis is conducted using OLS tool of ArcGIS selecting dengue cases as dependent variable and all other study parameters mentioned above as independent variables.

Results show that although a high R^2 of 0.747 (Adjusted R^2 = 0.733) value is found, the overall model has following problems.

a. Higher significance level (*P-values >0.05*) of regression coefficients including LST, population and population density.

b. An indicative of redundancy among variables is Large *Variance Inflation Factor* (VIF) (> 7.5). Two of the explanatory variables, population and built area, have VIFs greater than 7.5.

c. *Jarque-Bera Statistic* test is statistically significant ($p < 0.05$) that means that model predictions are biased and residuals are not normally distributed.

ii *OLS Regression between Dengue Cases and Built-up Area and Population Density*

Next, OLS regression is applied in order to examine the impact of *built-up area* and *population density* on dengue occurrences. Regression result shows that the *Adjusted R-Squared* value is 0.667, or 67%. This indicates that using *population density* and *built-up area*, the model is explaining 67% of the variation in dengue incidences. Also both explanatory variables are statistically significant but the *significance of the Jarque-Bera statistics is making this model biased and hence undesirable.* Also the *Koeker Test* is statistically significant that implies that the relationships between the dependent and some or all of the explanatory variables are non-stationary. This means, that the explanatory variable (*built-up area* and *population density*) might be an important predictor of dengue cases at some areas but may result in a weak prediction in other locations.

It is concluded, therefore, that study parameters might have spatially varying relationships and therefore, are not suitable for OLS global model. In the next sections, Geographical Weighted Regression (GWR) techniques are used to develop dengue predictive model. Geographically Weighted Regression analysis is a form of linear regression that can model spatially varying relationships between variables.

iii GWR between Dengue Cases and All Study Parameters

Geographically Weighted Regression (GWR) tool of ArcGIS is run selecting dengue cases as dependent variable and all other study parameters as explanatory variables. The model fails to execute and results cannot be computed successfully. As suggested by the model error message, this may be due to either severe global or severe local multicollinearity (redundancy among model explanatory variables). Different trail runs of the model are executed either by removing redundant variables from the model that have large *Variance Inflation Factor* (VIF-values > 7.5) in the previously run OLS models or trying other combinations of study parameters as explanatory variables. In the following section only that model is discussed in detail that is statistically strong as compared to other models tested in the study.

iv GWR for Dengue Cases, Built Area and Population Density

In this model two independent or explanatory variables; *built-up area* and *population density* are used to predict their influence on dengue outbreaks. The model results show that the *Adjusted R-Squared* value is 0.774 (R^2= 0.805). This indicates that using *population density* and *built-up area* as explanatory variables, the model can explain 77% of the variance in dengue incidences.

The GWR output is a standard residual map that is an indicator of model performance. Residuals are the portion of the total variability of the observed data that is unexplained by the model or the part of the model under and over predictions. In the standard residual map (Fig. 12) for the model developed between dengue cases and two explanatory variables – *built-up area* and *population density*, the red areas are under predictions where the actual number of dengue cases is higher than the model predicted values. The blue areas are over predictions where actual dengue cases are lower than predicted. A model is considered to be performing well when there is no clustered over/under prediction areas rather there is a random noise. Spatial clustering of over/under prediction areas is an indication of missing one or more key explanatory variables in the model. In Fig. 12, the standard residual map shows that the model is a little high at one place but a little low at some other place and there is no obvious structure of model over/under predictions. Spatial Autocorrelation tool (*Analyzing Patterns Toolset* of ArcGIS) can also be used on the model residuals to check if the residuals have a random spatial pattern or not. Whenever there is a structure/clustering of under/over predictions, the model is not trustworthy and it is an indication that some key explanatory variables are missing from the model. As shown in Fig. 13, the results of Spatial Autocorrelation analysis show that the regression residuals are randomly distributed since the *z-score* (=0.17) is not statistically significant. The null hypothesis of complete spatial randomness is, therefore, not rejected that ***confirms the randomness of the residuals required for a well performing model***.

Figs. 14 and 15, respectively, present the spatial distribution of regression coefficients of explanatory variables - *built-up area* and *population density*. Mapping these coefficients shows the relationship between each explanatory variable and the dependent variable that how they change across the study area. In Figs. 14 and 15, the red areas are locations where the explanatory variables, *built-up area* and *population density* respectively, are strong predictors of the number of dengue cases, whereas, the blue areas are locations where they are comparatively weak.

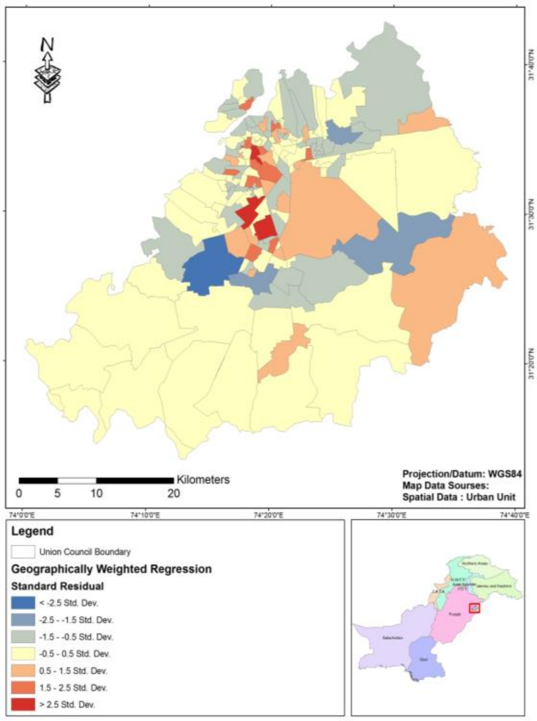
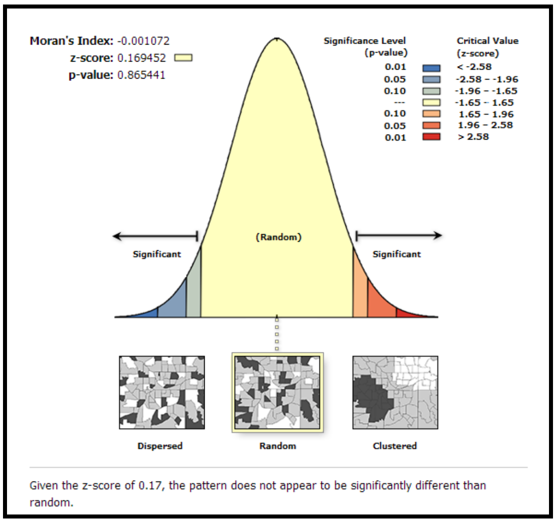


Figure 12. Standard Residual Map of GWR Model



Given the z-score of 0.17, the pattern does not appear to be significantly different than random.

Figure 13. Spatial Autocorrelation Report

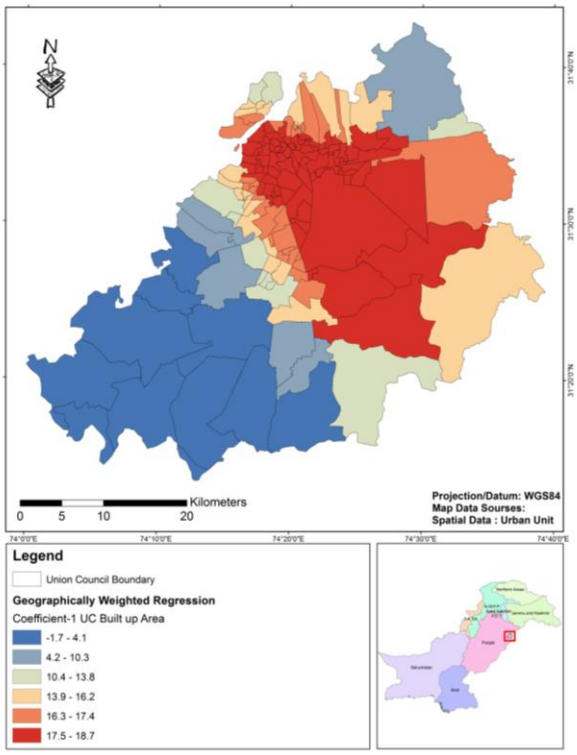


Figure 14. Spatial Distribution of Regression Coefficient - U.C. Built-up Area

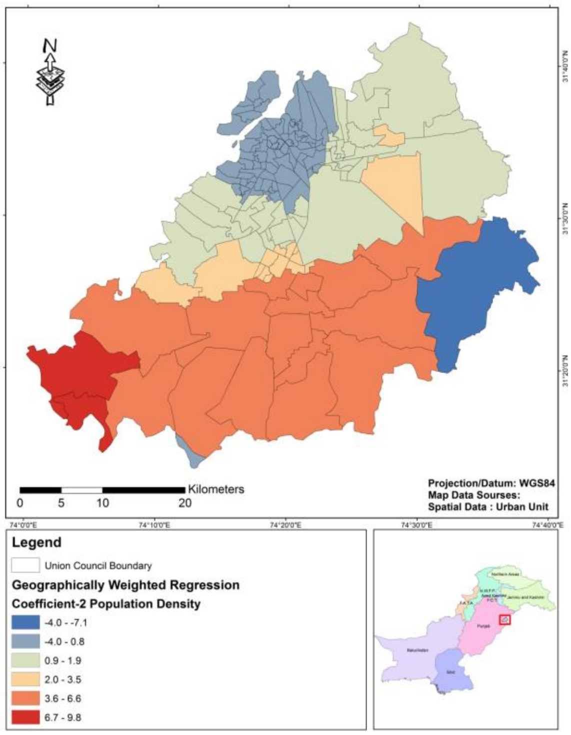


Figure 15. Spatial Distribution of Regression Coefficient – Population Density

C. *Dengue Prediction Map*

Geographical Weighted Regression model can also be used to predict values of dependent variables for locations within the study area with some projected values of explanatory variables. The prediction map is prepared assuming projected values of 10 % and 20 %, respectively, for *built-up areas* and *population densities* in all union councils. The predicted dengue cases are shown in Fig. 16 based on projected values of explanatory variables (*built-up area* and *population density*). There are some negative predicted values that are not acceptable and indicate model inaccuracy in these areas in giving acceptable estimations. Areas with negative values are shaded out and not included in analysis since these predictions are not reliable.

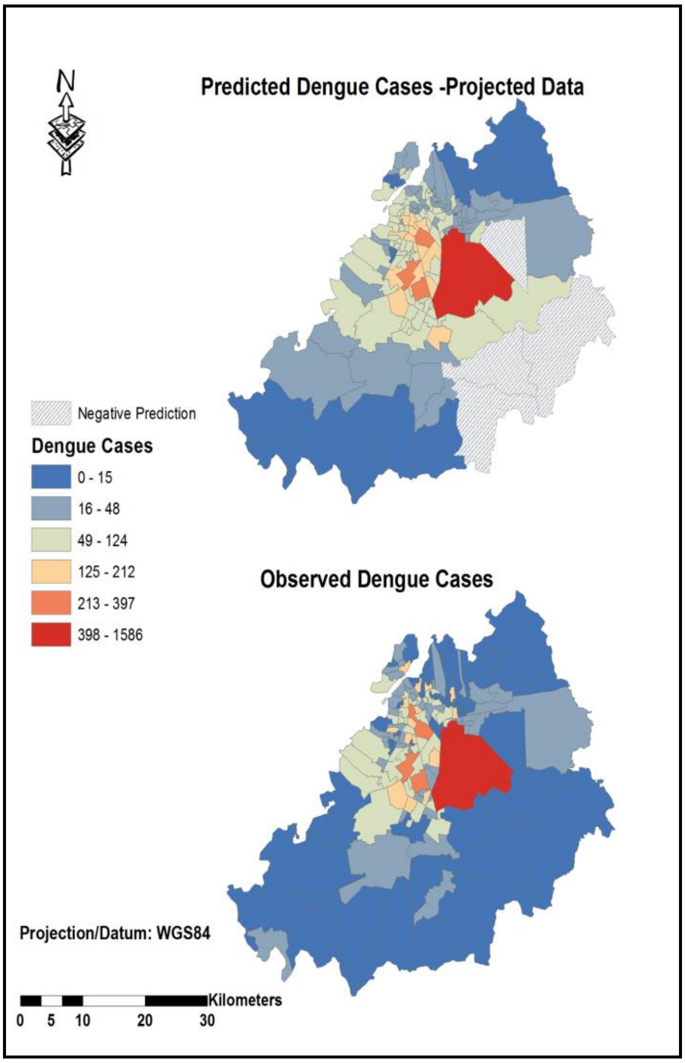


Figure 16. Dengue Prediction Map

V. Conclusion

The objective of this study is to develop a geo-statistical dengue risk map for Lahore city by identifying environmental, demographic and land-use/cover factors that significantly influence dengue outbreaks. This study analyzes the impact of these parameters on the distribution pattern of dengue outbreaks in Lahore city. The individual as well as synergistic impact of study parameters are evaluated and among them factors are selected that show significant influence on dengue outbreaks. Several factors including vegetation cover, land surface temperature, water, built-up area and population density are analyzed for their influence on the spread of the disease in the study area.

The model parameters used for this study are *Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), built-up area, population, population density* and *precipitation*. Linear regression models are built for sample UCs between dengue cases and rainfall temporal data. These models show statistically significant relationship between rainfall and dengue outbreaks in all sample union councils with R^2 ranging from 0.66 (minimum) to 0.8 (maximum). Ordinary Least Square (OLS) and Geographical weighted regression (GWR) analyses are employed to develop regression models between dengue cases and other study parameters.

Based on study results, it is concluded that study parameters are not suitable for OLS global model since no statistically strong model can be found using OLS. GWR analysis is a form of linear regression that can model spatially varying relationships between variables. The GWR model shows that using *population density* and *built-up area* as explanatory variables, the model can explain 77% of the variance in dengue incidences.

Densely populated and heavily built areas are the most vulnerable ones that may provide suitable breeding grounds for dengue virus to grow. There is a fear that these areas may experience severe dengue outbreaks if effective measures are not taken in order to control the disease or to minimize its risk. The geo-statistical dengue risk model developed in this study can be used to predict risk areas in the most vulnerable city Lahore that need special attention in order to effectively and efficiently manage and mitigate dengue outbreaks in future.

Acknowledgment

The authors would like to thank Member SAR, *Pakistan Space and Upper Atmospheric Research Commission (SUPARCO)*, Director *Pakistan Metrological Department (PMD)*, Director General (DG) *Punjab Disaster Management Authority (PDMA)* and Director *The Urban Unit* for providing data for this research work.

References

[1] WHO (2014). Global Alert and Response (GAR), Impact of Dengue. World Health Organization. http://www.who.int/csr/disease/dengue/impact/en/

[2] Ali, M., Wagatsuma, Y., Emch, M., F. Breiman, R. (2003). Use of a Geographic Information System for Defining Spatial Risk for Dengue Transmission in Bangladesh: Role for *Aedes Albopictus* In an Urban Outbreak. The American journal of tropical medicine and hygiene 69(06); pp. 634-640

[3] Anon. (2009). Distribution pattern of a dengue fever outbreak using GIS. Journal of Environmental Health Research 9(02).

[4] M. Umor, S., B. Mokhtar, M., Surip, N., Ahmad. A. (2007). Generating a Dengue Risk Map (DRM) Based on Environmental Factors Using Remote Sensing and GIS Technologies. Asian Conference on Remote Sensing.

[5] GOP (2009). Pre-investment studies for district Lahore. Directorate of Industries, Punjab. Government of Pakistan. http://punjab.gov.pk/sites/punjab.gov.pk/files/Lahore.pdfhttp://punjab.gov.pk/sites/punjab.gov.pk/files/Lahore.pdf