Dengue Hotspot Prediction in New Delhi

Shayon Mitra, Qazi Omair Ahmed

Abstract— This project discusses how to predict dengue fever cases and hotspots in a city based on population density, weather data and historical trend of cases. This can help the authorities concerned to take early action to prevent outbreak of dengue and prevent loss of lives and monetary loss. This uses machine learning techniques like random forest classifier and linear regression to predict the hotspots.

Keywords—machine learning; dengue fever; epidemic control; random forest

I. INTRODUCTION

Dengue fever is a common mosquito-borne viral disease worldwide and about half of the world's population is now at risk, especially in urban areas. It is a major epidemiology issue for India, where the national economic impact ranged around $1.1 billion.[1] The estimated number of dengue infections among individuals aged 5-45 in 2017 was 12,991,357.[2] Therefore, dengue fever is a vital issue for India regarding financial costs and the risk to public health. Prevention has therefore become important as there is no treatment drug for the disease. The purpose of this project is to predict dengue fever risk by using machine learning techniques using weather and urban data. Hence, there is a great potential for India to utilize data in order to generate insights for epidemic control, like dengue fever. In order to make this project usable, I have selected New Delhi for analysis.

II. LITERATURE REVIEW

A. *Concept of Dengue fever*

Understanding the nature of dengue fever is essential to address the spread, outbreak, and prevention of the disease. Dengue fever is a vector-borne disease, spread by a vector (Aedes mosquito) through biting a host (infected human).[3] Typically when a female mosquito takes a blood meal from an infected person, it takes two weeks of incubation period for the mosquito to be infectious to a healthy person. The disease can be transmitted from human to human, but the chance is low and not considered as the major cause of outbreaks.[4] Previous studies prove that temperature, precipitation, and humidity are critical to the mosquito life cycle.[5] Higher temperatures reduce the time required for the virus to replicate and disseminate in the mosquito.[6] Further, studies indicate that both dengue and severe dengue. geographical factors and climatic factors contribute to dengue fever outbreak, leading to the concept of landscape epidemiology. This concept emerges from the facts that most vectors, hosts and pathogens are usually associated with the landscape as environmental determinants.[7] Previously there are studies focusing on landscape characters contributing to mosquito-borne diseases. The World Health Organization (WHO) found that urbanization contributes to dengue fever outbreaks.[8] In recent years, increased studies have focused on landscape epidemiology using data mining and machine learning approaches for better disease prediction. Buczak A. et al. conclude that fuzzy association rule mining to extract relationships between clinical, meteorological, climatic and socio-political data is a valid method to predict dengue outbreak risk in Peru.[9] Sarfraz M. et al. utilized satellite images and Decision Tree methods to predict mosquito habitats in Thailand and conclude that the most influential factors are temperature, humidity, rainfall, population density, elevation and land cover.[10] Yong-Su Kwong et al. used machine learning techniques such as Principal Component Analysis, Support Vector Machine and Random Forest to indicate that land use patterns along with temperature are important predictors for the occurrence of mosquitoes in South Korea.[11] There are also researchers focusing on the influence of mobility on urban epidemiology. Stolerman L. M. et al. studied network across neighborhoods within cities and the movement of residents from home to work place or place for general activities. The generated directed network, with people's movement as weights of edges, allows predicting possible outbreaks by approximating the reproduction number of dengue fever in Rio de Janeiro.[12]

B. *Data Sources*

Based on previous research and our own investigation, three major data sources are chosen:

I. Disease Cases Data

The National Vector Borne Disease Control Programme website includes the total number of dengue fever cases nationally and statewise. The dataset holds information on all infectious diseases recorded. The Dengue Cases Data, on the other hand, is provided by the National Environment Agency and includes data of dengue cases on a daily basis, together with the geographical location of active cases as point data.

II. Weather Data

In order to investigate a possible relationship for New Delhi, historical weather data is matched with disease cases data based on the week of the year, while weather data is lagged by two weeks to account for the mosquito's life cycle.[13]

III. Urban Data

The city officials of New Delhi provide datasets from quite a few public agencies, which include urban data on environment, demography, infrastructure, transportation and facilities.

1) Street network will act as a proxy for the transportation network across New Delhi and together with the bus stops will act as an approximation of the movement of people. Lot density will act as an indirect proxy for population density, since population information is not available on a granular level.

2) Trash bins, parks and water facilities willqu be used to account for artificial water bodies and probable mosquito habitats.

3) Mosquito habitats identified by New Delhi's Environmental Agency will be used to evaluate whether the above mentioned habitat proxies are relevant and if not, to supplement the analysis.

III. METHODOLOGY

The analysis is divided into two parts to capture temporal and spatial patterns of dengue cases. The first analysis, a temporal analysis, aims at investigating the correlation between meteorological data and dengue fever cases based upon national historical data, whereas the second analysis, a spatial analysis, focuses on geographical and demographic features of New Delhi as potential predictors. The two analyses are separated.

A. Temporal Data Analysis

For the temporal analysis, disease cases (dependent variable) and weather data (independent variable) are merged based on the week of the year into a final dataset, where weather data is lagged by two weeks due to the virus' incubation period. Weather data includes multiple features including temperature, humidity, wind speed and rainfall. In order for comparisons across weeks and define outbreaks, weekly cases are binned to create five categorical labels. Upon this final dataset, a cross-validation is applied in order to properly train, validate and test/compare both Linear Regression (OLS) and Random Forest Classifier (RF). OLS was used to for establishing a comparable basis and judge on the significance of each feature included. RF was employed to possibly improve accuracy given the characteristics of the data, i.e. classification problem with weak linear relationships.

B. Spatial Analysis

The spatial analysis focuses on spatial and environmental data in order to predict disease cases across New Delhi. Several features are used as predictors, including population density, geographical characteristics (land cover, waterbodies, lot density), community facilities (hospitals, parks, aquatic facilities) and urban infrastructure (street network, bus stops, taxi stops). The outcome variable has binary labels, identifying those areas with presence of dengue cases.

C. Machine Learning Details

In order to form the final dataset for machine learning models, it is necessary to conduct data processing techniques to integrate multiple layers of spatial data and ensure that multiple features are aligned correctly regarding their respective geo-location. This study utilizes raster analysis in ArcGIS that converts multiple vector shapefiles into layers of raster images with same resolution and image size. Specifically, the Kernel Density function converts point data into a raster image, which indicates the density of points based on the quartic kernel function. Kernel density estimation calculates the spatial distribution of cumulative incidence per sq. km and it is a common technique for spatial analysis on weighted density over a gridded surface.[14] By properly setting image extent, cell size, resolution and pixel position, such raster transformation converts multiple datasets into gray scale images (0-255) with same resolution (1 sq. km pixel size). The resulting image size is 850*250 pixels. The final dataset is a data frame, where each observation represents a pixel and each column a spatial or demographic feature of that pixel. Next, Logistic Regression and Random Forest Classifier are used for predictive modeling. Logistic regression predicts the probability of having a dengue case in each area and allows for determining the relative importance of the included features. This model is used to derive a general sense of the data and the significance of the model, judged additionally by the Area Under Curve (AUC) score. Since labeled data represents approximately 1.2% of the data, the AUC's characteristic of giving the proportion of the time the guessed label equals the actual label makes this metric better than the accuracy score, since it is less affected by sample balance. Second, RF is employed to crosscheck findings from logistic regression and make use of a more sophisticated model, accounting for nonlinearity of the data. The parameters of the RF are taken from cross-validation, taking the best performing set (highest AUC score) among different combinations of possible parameters (using K-fold split of 6),[16] including the number of estimators or trees in the forest between 100 and 1000, the number of features considered for the best split among and log, and the function to measure the inequality of split among Gini impurity and information gain.[16]

IV. FINDINGS AND CONCLUSION

A. Temporal Analysis

Analysis on historical weather data and dengue cases data shows that temperature is the most correlated feature compared to other features such as rainfall and wind speed. Rainfall and wind speed are more sporadic and volatile, making it hard to find a suggestive pattern. Indeed, the correlation matrix confirms the intuition drawn from the figures. Scatterplots of each of the features against the independent variable is contained in the appendix Figure III. Linear Regression confirms that temperature and rainfall are the only significant features with a p-value of 0.000 and 0.03, respectively, being associated with dengue fever positively. The R-squared value reflects the findings, with a low value of 8%. RF performs relatively better with an out of sample accuracy score of 34%, with the following parameters: 300 trees in the forest, a split of log features, and entropy as inequality measure. Hence, using just meteorological data is not sufficient to accurately predict dengue fever.

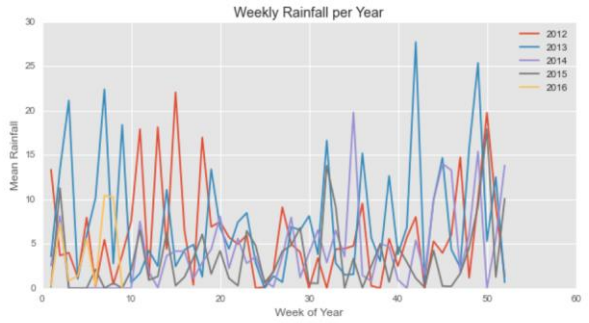
B. *Spatial Analysis*

Mosquito habitat correlation with the dependent variable is suspected due to the fact that such data is collected by the same agency as dengue fever cases and it is highly likely that New Delhi's authorities identify mosquito habitat based on reported cases. Excluding this predictor, the AUC score was calculated to be 50%, which means that the algorithm is performing with the same accuracy as random guessing. However, pseudo R-square is approximately 20%, which suggests some predictive power. In addition, all included features are statistically significant at an alpha level of 5%. The marginal effects, i.e the change in probability of the dependent variable given changes in the independent ones, are the strongest in positive terms for transportation related variables (street network and bus stops) as well as trash bins. Parks and the total population have a negative effect. This suggests that higher mobility contributes to higher dengue fever risks. The association of population is harder to reason, because lot density and population do not fully capture population density and building density.

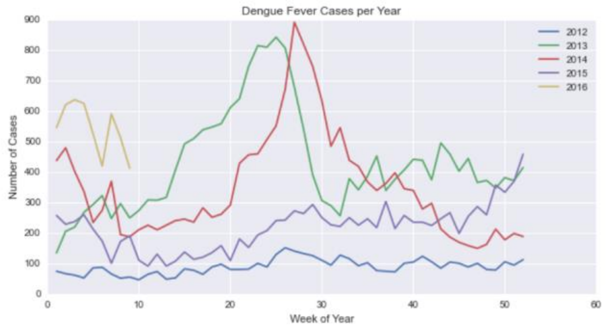
C. *Conclusion*

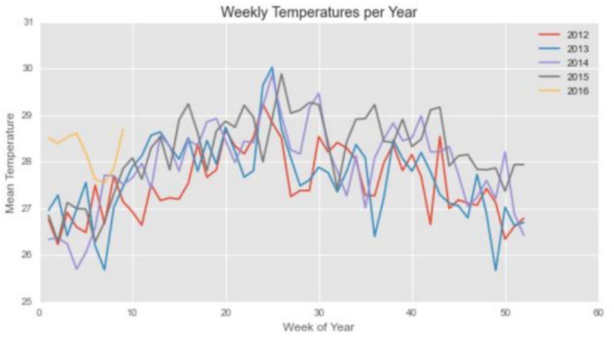
This study demonstrates the potential application of machine learning with urban data for dengue fever prediction in urban settings. Logistical Regression and Random Forest model are both a potential approach, and the results indicate that Random Forest performs better in terms of AUC score. The current model does not integrate both temporal and spatial data due to limitations in data, stemming from granularity and availability. Hence, this paper was not able to integrate the low predictive power of meteorological data into the spatial analysis in order to create a better approximation of landscape epidemiology. Further, variables used for evaluating transportation networks are highly static and only approximate. Thus, real time data from Google maps, Twitter or even street cameras could potentially increase the predictive power. Ideally, data capturing work-home dynamic should be included in future work. This paper's model was cross-validated using train and test sets of one two-week dataset in order to reduce over fitting and improve AUC scores. Future analysis should validate findings on future data published by New Delhi's authorities. Through this, predicted areas could be cross-referenced with newly erupted cases, checking whether the model predicts correctly. Nevertheless, the model in this paper has a good mathematical performance. Addressing the shortcomings and limitations, next steps should aim at investigating the RF model's false negatives. This should ultimately improve the usefulness of this paper's analysis in terms of managing New Delhi's(and therefore India) financial and human resources when combating dengue fever cases.

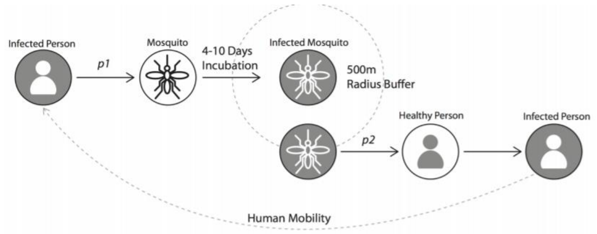
GRAPHS



```
plt.savefig('dengue.png')
```







REFERENCES

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4257651/

[2] https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(19)30249-9/

[3] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Chen L. H. , Wilson M.E. (Oct. 2010) Dengue and chikungunya infections in travelers. Current Opinion in Infectious Diseases 23(5): 438-44

[5] Mosquito Life Cycle, Singapore Centers for Disease Control and Prevention. Retrive from: http://www.cdc.gov/dengue/entomologyEcology/m_lifecycle.htm

[6] http://www.cdc.gov/Dengue/entomologyEcology/climate.html

[7] Pavlovsky, E.N. (1966) Natural Nidality of Transmissible Diseases, With Special Reference to the Landscape Epidemiology of Zooanthroponse. Urbana, III.

[8] Gubler D. Dengue, Urbanization and Globalization: The Uhnoly Trinity of the 21st Century. Tropical Medicine and Health 2011; 39(4) Retrieve from: https://www.jstage.jst.go.jp/article/tmh/39/4SUPPLEMENT/39_2011-S05/_article

[9] Buczak A., Koshute P., Babin S, Feighner B. and Lewis S. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Medical Informatics and Decision Making. 12: 124 (2012)

[10] Sarfraz M., Tripathi N., Faruque F., Bajwa U., Kitamoto A. and Souris M. Mapping urban and peri-urban breeding habitats of Adedes mosquitoes using a fuzzy analytical hierarchical process based on climatic and physical parameters. Geospatial Health 8(3) 2014. pp. 685-697

[11] Kwon, Yong-Su, Mi-Jung Bae, Namil Chung, Yeo-Rang Lee, Suntae Hwang, Sang-Ae Kim, Young Choi, and Young-Seuk Park. "Modeling Occurrence of Urban Mosquitos Based on Land Use Types and Meteorological Factors in Korea." International Journal of Environmental Research and Public Health IJERPH 12, no. 10 (2015): 13131-3147

[12] Stolerman, Lucas M., Daniel Coombs, and Stefanella Boatto. "SIR-Network Model and Its Application to Dengue Fever." SIAM J. Appl. Math. SIAM Journal on Applied Mathematics 75, no. 6 (2015): 2581-609

[13] https://www.wunderground.com/history/monthly/in/new-delhi

[14] Statistical methods on Kernel Density Estimation. Human Cutaneous Anthrax, Georgia. Technical Appendix (2010-2012) Centers for Disease Control and Prevention. Retrieve from: http://wwwnc.cdc.gov/eid/article/20/2/13-0522-techapp1.pdf

[15] http://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it 17

[16] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html