Analysis and Prediction of Dengue Disease Spread

Siddhanth M
Department of CSE
PES University
Bengaluru, India
siddharththe8055@gmail.com

Uthpal P
Department of CSE
PES University
Bengaluru, India
uthpaall@gmail.com

Vishwas R
Department of CSE
PES University
Bengaluru, India
vishwasr6543@gmail.com

***Abstract*—The primary goal of this project is to use statistical analysis methods and machine learning techniques for the analysis, prediction and forecasting of the spread of dengue disease to gain insights into the numerous factors that are associated with it and how they influence the spread of dengue disease. Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics, which often change in time and space. It is well known that its disease dynamics are highly influenced by multiple factors including population susceptibility to infection as well as by microclimates: small-area climatic conditions which create environments favourable for the breeding and survival of mosquitoes. Additionally socioeconomic factors and population demographics also contribute to the spread of dengue. Here, we plan to present a novel machine learning dengue forecasting approach, which, dynamically in time and space, identifies local patterns in weather and population susceptibility to make epidemic predictions.**

***Index Terms*—Dengue Fever (DF), Time Series, Boosted Regression Trees, Generalized Additive Model(GAM), Diurnal temperature range (DTR)**

I. Introduction

Dengue disease is a deadly viral disease caused by mosquitoes. Since the spreading of dengue is done by mosquitoes, many studies show that the spreading of this disease is positively correlated with climate factors such as temperature, humidity and precipitation amount. With having enough historical data related to the dengue disease and the climate changes, it invites the question of whether it is possible to use the correlation of the data to predict an epidemic before it happens. Although it is possible to study the data in conventional statistical methods, there is an emerging trend of using data science for analysing and exploring large sets of data related to the area of disease forecasting.

II. Literature Review/ Related Work

A. Paper 1

Environmental factors such as temperature, precipitation, humidity and wind play a key role in mosquitos' ability to breed successfully resulting in widespread mosquito-borne diseases. In the paper [1] the researchers analyzed the correlation between these meteorological factors and the number of daily cases obtained in Guangzhou, China for a period of six years. Correlation analysis was done on the environmental factors to find a significant amount of correlation, hence these factors were fed in separately to different models and based on the performance of these models (using AUC) only the unique predictors found were chosen for further analysis. The Boosted Regression Tree (BRT) model was applied. A combination of algorithms that used recursive binary splits to choose predictors having the least amount of entropy used in the regression tree model was boosted such that the BRT model built a large ensemble of small regression trees to improve predictive performance. A 10-fold cross-validation technique that stratified the data by prevalence was used to obtain the optimal model with the least residual deviance. Then the optimal lag between the environmental changes and the record of the cases arising from these changes was obtained to be 58-60 days by running the model considering the lags in the range of 0 - 120 days and selecting the lag period having the highest AUC. Although this model provided a binary prediction for dengue fever risk, it was hard to distinguish between "significant outbreaks" and "random noise". We could address this by incorporating more continuous case data, and applying different statistical distribution model.

B. Paper 2

In paper [2] the authors used time-series regression to obtain the hierarchical relationship between timeliness of Dengue Fever (DF) surveillance systems, mosquito density, imported cases, meteorological factors and autochthonous DF incidence rates. Rather than considering the Mean temperature of each day as done in paper [1], this model took the Diurnal Temperature Range ("Max temp - Min temp" of a single day) into consideration for the changes in DF incidence rates. The fact that short-term temperature fluctuations could substantially alter the incubation period for parasites was brought into consideration with the use of a calculated diurnal temperature range.

The Classification And Regression Tree (CART) model was used as it is considered as a non-parametric flexible model which built using ensembled learning will express the nonlinear relationships, high-order interactions, and missing values too. Two measures were used for regression trees: least squares (Gini splitting rules) and least absolute deviations (two splitting rules). Large trees were obtained and were cross-validated to prune and obtain the best-trimmed tree on the basis of the cost complexity function:

$$CC(T) = err(T) + \alpha L(T) \tag{1}$$

where err(T) is the misclassification error of the tree T, L(T) is a measure of complexity of the tree, either length or the number of nodes in the tree and alpha is the cost-complexity parameter which penalizes for the overcomplexity of the tree. This parameter alpha is varied through the learning and testing samples to obtain the minimum cost complexity tree. This study used the lagged moving average (MA) value to calculate the lagged effect on the series of factors.

These papers [1] and [2] suggested that the Regression tree models provide a non-parametric approach that can potentially better accommodate these complex interactions since they avoid some of the assumptions associated with linear regression.

C. Paper 3

The goal of the research paper [3] was to create a system that can use the available relevant information about the factors responsible for the spread of dengue and; use it to predict the occurrence of dengue within a geographical region, so that public health experts can prepare for, manage and control the epidemic. It presents a machine learning-based methodology capable of providing forecast estimates of dengue prediction in each of the fifty districts of Thailand by leveraging data from multiple data sources.

The paper shows an increase in prediction accuracy of the model with an optimal combination of predictors which include: meteorological data, clinical data, lag variables of disease surveillance, socioeconomic data and the data encoding spatial dependence on dengue transmission. It uses Generalized Additive Models (GAMs) to fit the relationships between the predictors (with a lag of one month) and the clinical data of Dengue hemorrhagic fever (DHF) using the data from 2008 to 2012. Using the data from 2013 to 2015 and a comparative set of prediction models, it evaluates the predictive ability of the fitted models according to RMSE and SRMSE as well as using adjusted R-squared value.

According to the paper the factors associated with dengue cases are:

- Meteorological (DTR- Diurnal temperature range-The difference between max temperature and min temperature of a day and rainfall) and socioeconomic data (monthly garbage collection in each district) to the time series of dengue incidences in that particular district.
- Dengue transmission in a specific month in a district with the data of past occurrences of dengue in its surrounding districts.

Using the above mentioned predictors the authors come up with various models like

1. Meteorology Optimal model (includes lag of 0-4 months for DTR and mean monthly rainfall)

2. Optimal Lag Surveillance Model (includes lagged dengue count data for 1, 2 and 23rd months)

3. Optimal Meteorology and Lag Surveillance Model (includes lagged meteorology data for 1-3 months and lagged dengue count data for 1, 2 and 23rd months)

4. Optimal Representation Model Combination of (3) with lagged dengue count data of surrounding districts for 1st and 2nd months.

5. Social-economic data Included Combination of (4) with garbage collection data of each district as the social capital.

The best model for the dengue prediction is the one that includes lagged meteorological data (rainfall, DTR), lagged dengue data of the target districts as well as their surroundings and the socioeconomic data. The model is observed to have lowest errors (RMSE, SRMSE) and have the best fit (measured in R-sq.adj)

A number of limitations are apparent in this research. First, the predictive model that gives the best results, can explain only 73 percent of the variation in the occurrence of dengue cases. The remaining 27 percent unexplained variation could be due to the influence of other factors. The out-of-sample predictive performance was considerably worse than that of in-sample performance. The authors point out that rather than dismissing it as a case of overfitting, the case demands that we look into the facts and plausible causes behind it.

All in all, it can be viewed as a foundation for our analysis, giving a brief idea as to which factors influence spread of dengue in order for us to improve, develop further and newer insights.

D. paper 4

The research paper [4] presented a novel approach to forecasting dengue fever outbreak in Brazil using a single, dynamic and flexible modelling framework that uses only two weather variables and historical information on yearly dengue activity. The approach automatically learns from weather and population susceptibility patterns of any inputted yearly time series of dengue incidence and leverages the best historical predictions to generate an ensemble forecast. Importantly, the models can be immediately extended to other locations, requiring no location-specific manipulation or inputs aside from a globally available time series of daily temperature and precipitation as well as a complete yearly record of dengue incidence.

The modelling framework consists of :
Signal preprocessing: Defining time intervals of varying sizes for a time series of weather data.
Time-series feature extraction: Extracting a simple summary measure for two weather variables with known influence on mosquito-borne disease dynamics. Eg: temperature and frequency of precipitation.
Independent model training and prediction: Training a collection of independent SVM classifiers on historical information from each unique time interval, and generating an out-of-sample epidemic prediction for the following year.
Model selection: Choosing the best models, representing strongly predictive periods of the year preceding outbreaks, based on historical out-of-sample prediction accuracy and out-of-sample performance of neighbouring time intervals.

Ensemble prediction: Determining a final out-of-sample epidemic forecast by a majority vote of the selected top models.

Dengue cycles: Implementing a decision rule governed by the second and third-order Markov transition probabilities, reflecting the transition between consecutive sequences of epidemic and non-epidemic states.

The authors observed that, in general, the probabilities were moderately calibrated, i.e. roughly 80% of predictions made with 0.8 probability were epidemics; however, the small sample size (i.e. six out-of-sample years for each of the 20 cities) limits the ability to interpret this feature appropriately. It was found that this measure of separability was not a particularly good indicator of accuracy; that is, the approach failed even in scenarios with high separability. Several factors may be driving this finding, including insufficient training data and the influence of factors beyond weather (e.g. sociodemographic characteristics, land use) on outbreaks.

III. Data and Problem Description

In 2015, the National Oceanic and Atmospheric Administration of US as a part of the dengue forecasting project provided the time series data containing the number of dengue cases and the atomspheric conditions at that time of the cities of San Juan,Puerto Rico and Iquitos,Peru during the period of 2001-2009. This data is provided as a part of a challenge problem in drivendata.org in order to predict the number of dengue cases that would arise per week in the future in the 2 cities by analysing the historical data provided. The primary goal of this project is to use statistical analysis methods and machine learning techniques to derive insights from the data provided and solve the above challenge with utmost accuracy and precision.

IV. Exploratory Data Analysis

The data given was first split into two different sets on the basis of the two different cities San Juan and Iquitos. Then the missing numerical values present in these datasets were replaced with their column averages and the rows with missing descriptive data were dropped. The data which is less than 0.05 percentile and greater than 0.95 percentile are identified as outliers and are removed. To find the initial insights on the data present, correlation analysis was done by plotting the correlation heat map using Pearson's Correlation Coefficient between the continuous random variables present in the datasets belonging to both the cities as seen in fig.1 and fig.2. The plot on the aggregated number of cases per week of each year peaked around the 40th weak of the year [fig.3], implying that the cases were the highest in the monsoon season. We can also observe occasional peaks in the number of cases between every few years [fig.4]. The number of cases in Iquitos were less compared to San Juan, this is because of the population differences between the two cities [fig.5].

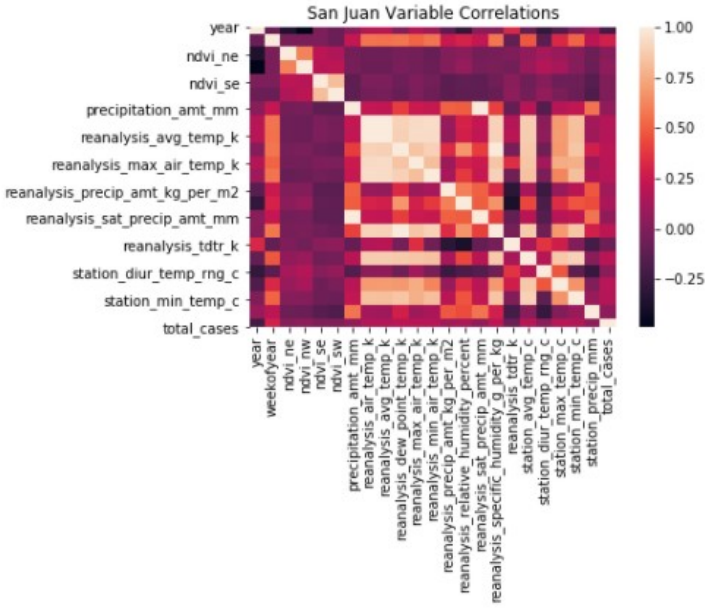

Fig. 1. San Juan Correlation Heat Map

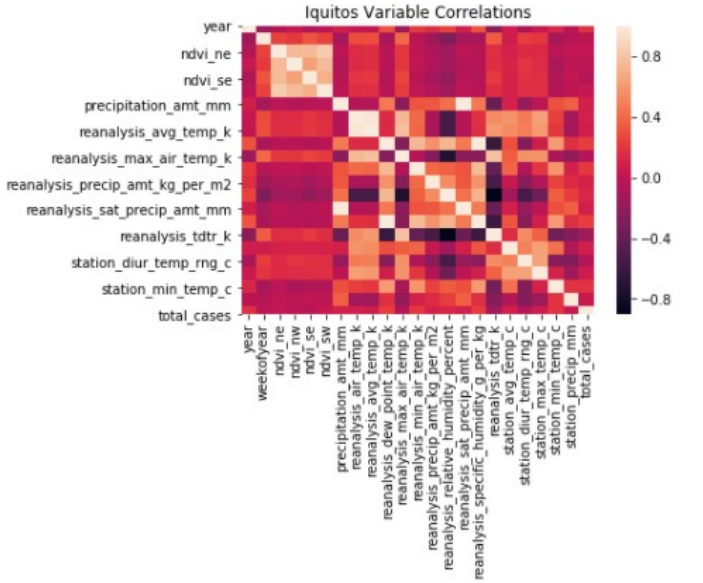

Fig. 2. Iquitos Correlation Heat Map

References

[1] H. Gu, R. Leung, Q. Jing, W. Zhang, Z. Yang, J. Lu, Y. Hao, and D. Zhang, "Meteorological Factors for Dengue Fever Control and Prevention in South China," International Journal of Environmental Research and Public Health, vol. 13, no. 9, p. 867, Aug. 2016 [Online]. Available: http://dx.doi.org/10.3390/ijerph13090867

[2] K.-K. Liu, T. Wang, X.-D. Huang, G.-L. Wang, Y. Xia, Y.-T. Zhang, Q.-L. Jing, J.-W. Huang, X.-X. Liu, J.-H. Lu, and W.-B. Hu, "Risk assessment of dengue fever in Zhongshan, China: a time-series regression tree analysis," Epidemiology and Infection, vol. 145, no. 3, pp. 451–461, 2017.

[3] Jain, R., Sontisirikit, S., Iamsirithaworn, S. et al. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect Dis 19, 272 (2019). https://doi.org/10.1186/s12879-019-3874-x

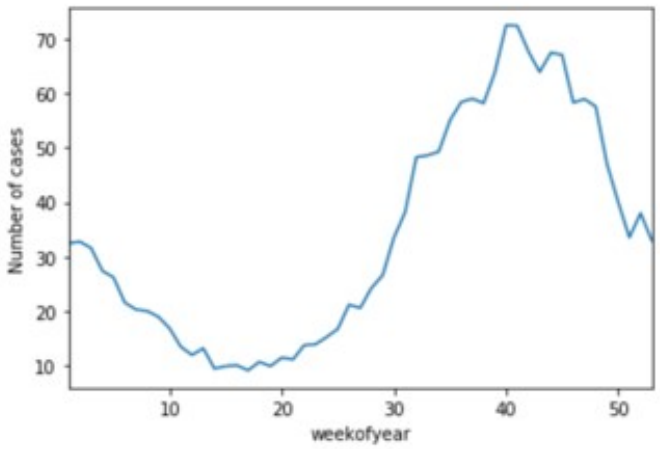

Fig. 3. Seasonal impact on number of cases

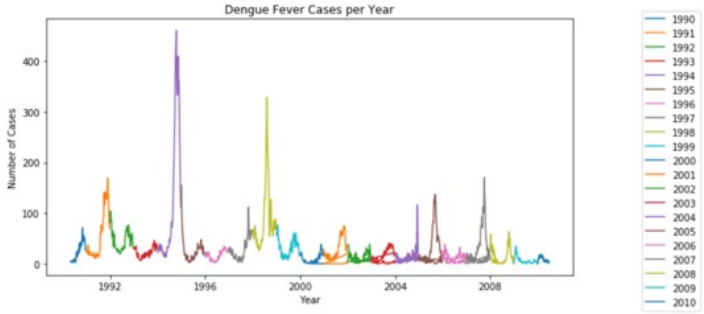

Fig. 4. Dengue fever cases per Year

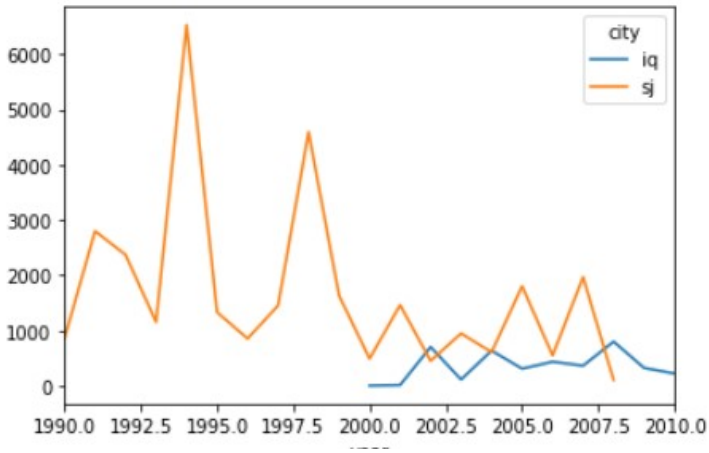

Fig. 5. Comparision of number of cases between the two cities

[4] McGough, Sarah F. and Clemente, Leonardo and Kutz, J. Nathan and Santillana, Mauricio. A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles Journal of The Royal Society Interface 18,179 (2021) doi:10.1098/rsif.2020.1006