Real-Time Disease Forecasting using Climatic Factors: Supervised Analytical Methodology

Garima Makkar
Senior Business Analyst,
Tata Consultancy Services,
Bangalore,India
garima.makkar@tcs.com

***Abstract*— Weather being an uncontrollable factor, often has direct effects on human mortality rates, physical health, mental injury and other health outcomes. Extreme climate incidences and gradual changes in weather are making us more vulnerable to disease outbreaks. In general, there are three ways by which variation in climate affects such diseases: by affecting the virus, the vector or host and spread of a disease. According to 1996 World health organization (WHO) report, 30 new diseases have come into existence in the past 20 years. Additionally, there has been a re-emergence and redistribution of various arthropod-borne diseases such as dengue, malaria etc. on a global scale. Events like rainfall, humidity, temperature etc. have well-defined role in the transference cycle. Any changes in these events can lead to increase in incidence of these diseases.**

The worldwide pandemic about abroviral diseases demands the need for developing early warning system (EWS) for infectious diseases by considering climate change. Past studies incorporates only the historical weather statistics into account. However because of increasing uncertainty and climate variability, the traditional systems in this context are getting outstripped. In this paper, we'll propose our methodology for predicting number of dengue cases that are likely to occur in real time on the basis of five-day weather forecast. Our analysis is applicable globally and enables comprehensive scenarios of daily disease outbreaks to be explored using real-time weather API, preparing society against any health related risks arising due to variability in climate.

Keywords*— *Weather, Climate, Disease, Dengue, Real-Time, Early warning system

I. Introduction

According to 1996 World health organization (WHO) report, 30 new diseases have come into existence in the past 20 years. Additionally, there has been a re-emergence and redistribution of various arthropod-borne diseases such as dengue, malaria etc. on a global scale. One of the major reason behind this variation in the incidence of mosquito-borne diseases is extreme climate events and changes in weather patterns. A virus, a vector and living environment are the three crucial components for most infectious diseases. And a suitable set of weather conditions are required for their survival, breeding, distribution and spread of these components. So any changes in climate and weather patterns may affect the vector-borne diseases by impacting the pathogens, vectors and their transmission environment.

Extreme climate incidences and gradual changes in weather are making us more vulnerable to disease outbreaks such as dengue, which is considered as the most widespread abroviral disease in humans around the world. According to WHO Report, every year there are around 50-100 million dengue incidences, and more than 500,000 cases are hospitalized. One of the important reasons behind the dengue transmission is climate change. Such vector-borne diseases (like malaria, dengue, lyme disease etc.) are quite sensitive towards meteorological variables such as precipitation, temperature and humidity. For instance, warm temperature is important to gonotrophic cycle and feeding behavior of adult dengue vectors, as well as the rate of viral simulation and speed of larval development. Similarly, rainfall-induced standing water is considered critical for dengue vectors to breed. Thus all in all it can be said that events like rainfall, temperature, wind speed etc. have well defined role in the transference cycle and any change in these events can lead to increase in the incidence of this disease.

The impact of global epidemic on mankind facilitates the need for developing early warning system (EWS) on infectious disorders with respect to the climate change. Past studies incorporates only the historical weather statistics into account. However because of increasing uncertainty and climate variability, the traditional systems in this context are getting outstripped. Also, weather prediction which is an attempt by meteorologists to forecast the state of atmospheric parameters such as humidity, temperature, rainfall etc., is considered as one of the most complex task in today's world. Understanding how factors like weather and climate affects the disease occurrence in a specific geographic region is an eternal part of disease forecasting. Hence, models based on weather data helps to predict where and when the human cases are most likely to occur. Such information assists to target the control resources and restricted prevention and may finally decrease the burden of diseases.

Keeping this in mind, we propose a Real-time Disease Prediction framework that analyses the five-day weather forecast data being extracted using an API key and provides a number of dengue cases that are likely to occur in these upcoming five days. Given the necessary weather data, we have used a supervised machine learning algorithm called Random Forest to carry out this real-time disease prediction analysis. Our analysis as explained in this paper is divided into following sections. The Section 2 discusses some of the past works done in context of disease and weather patterns. Section 3 gives a brief about our problem statement while Section 4 explains our proposed methodology. The result and conclusion are explained in Section 5, 6 and 7.

II. LITERATURE REVIEW

This section gives a brief about some of the research works that have been done in the context of weather and disease so far. Vector-borne diseases have been reported at high elevations in various parts of the world and thus are growing public health threat these days. Arthropods being sensitive to climatic factors have led various researchers to explore this area of study in detail. Basically, these studies can be divided into two groups: 1) Theoretical work that describes the nature of the impact and 2) Empirical work which explains the impact of weather variation on human health. The following is the description of few of these works explaining the impact and magnitude of climate variability on human population.

Many researchers have explored the responsiveness of mosquito cases and its transmission patterns with respect to changes in meteorological variables. For instance, [1] &[2] suggested that the viral development and its transmission occur more frequently and more rapidly at warmer temperatures. Reference [3] applied forecasting system to predict the malaria risk in Botswana which initiated timely anticipated mitigations and thus helped malaria decision-makers to use the results for improved resource allocation. Another consensus suggests that when the temperature is between 27°C & 30°C then the spread of virus is at its peak [4].Reference [5] presented a thermodynamic model showing how daily temperature fluctuations affect the vector pathogen interactions and why short-term variations in temperature are important when studying the disease transmission dynamics. Also, various researchers studied the health effects of global warming and disrupted climate patterns in detail and concluded that the long term climate warming tends to impact the geographic growth of various infectious diseases by creating the opportunities for more clustered disease outbreaks to take place[6][7][8].

Few empirical models have also been developed for estimating the weather effects on different infectious diseases. Disease prediction using time series analysis has been explored by many. For example, Reference [9] studied the relationship between climatic variables and human plaque incidences using Poisson regression and concluded that the variations in plaque risk can be estimated by temperature and time-lagged amount of late winter precipitation. Similar to this work, a time series Poisson multivariate regression model to predict dengue cases in Singapore over the period 2000-2010 has also been performed. Seasonality, autoregression, trend and various lag times were considered in their analysis to find the optimal dengue forecasting period using cumulative rainfall and weekly mean temperature as the only independent variables [10]. Reference [11] [12][13],all have proposed similar time-series framework analysis for various other vector-borne diseases and in different geographic regions. Some of the latest works include References [14] [15] etc. have given systematic view on the effects of weather and climate on particular category of diseases and recommended the factors important for framing the health policies.

Disease prediction using Support Vector Machine (SVM) has also been performed by few of the researchers. Reference [16] employed Support Vector Machine regression to forecast the number of disease cases based on the climatic factors. Their methodology resulted in a strong correlation between the monsoon seasonality and dengue virus transmission. References [17] [18] [19] etc. are some of the other disease forecasting works reporting varying associations and lagged effects between weather and disease cases.

Thus, there has been limited set of recent studies which developed site-specific multivariate regression models using different combinations of weather variables to predict different vector-borne diseases (e.g., malaria, dengue etc.) in different regions of the world. But almost all of them are based on historical weather data and none of them forecast these incidences in real time. While our analysis takes into account both the historical as well as future weather data where the disease forecasting has been done in a unique manner.

III. PROBLEM STATEMENT

The objective here is to create an Early Warning system that will predict the disease incidences (in our case it is dengue incidences) in real time given the five-day weather forecast. Since weather data is complex, non-linear in nature so the traditional methods aren’t effective and efficient to solve such problems. The proposed methodology evaluates the developed models by tuning different combinations of parameters to predict the disease (in our case it is Dengue) in real time given the five day weather forecast .The criteria used for model selection is Root mean square error. Contrary to similar research work, the data model and methodology suggested in this paper resulted in higher accuracy and better performance (i.e. reduced computational complexity). Thus, this analysis focuses on weather as the fundamental factor behind dengue epidemics which may allow us to reduce the timeframe of high risk dengue infection.

IV. METHODOLOGY

This section explains the methodology that is being followed to carry out the necessary experimentation for our analysis. The first part describes the dataset which is being utilized, followed by the steps performed to predict the dengue cases on the basis of weather conditions in real-time.

A. Data

Our main purpose is to predict the occurrence of dengue cases that are likely to occur in Philippines using its current and future weather conditions. For this purpose, we have used the data which is a merger of two different datasets: - 1) Dengue cases with respect to historical weather conditions and 2) five days forecasted weather data based on API key. Former dataset has been taken from Kaggle which is the predictive analytics competition platform while the latter is from Openweathermap, a service providing historical, current and future weather conditions for each and every city based on the API key. Using this amalgamation of datasets, we performed the following steps to build our predictive model

B. EDA

Exploratory data analysis is considered as the first and foremost step in any data analysis procedure. Basically, this approach employs a set of techniques to tackle various tasks such as detection of outliers, spotting missing data, maximizing insights in a dataset, uncovering underlying relationships etc. Most of these techniques are graphical in nature with a few quantitative techniques, helping to display the data to speak for itself. And one of the best ways for data analysts to present their analysis outside the industry is through visualisations. Thus, keeping this is mind, we also performed EDA so as to see the relationship or correlation (if any) present between the variables/features in our raw dataset. For example, we plotted scatter plots to see the relationship/correlation between: 1) number of dengue cases versus temperature mean and 2) number of dengue cases versus humidity. From these we got to know how dengue number varies with different weather elements. Similarly, a bar plot (as shown in Fig. 1) is being plotted to see the monthly dengue cases in Philippines for consecutive two years. The stacked bar-chart showcased here not only allow us to see the yearly totals, but also get a rough yet helpful understanding of the number of dengue cases occurred within each month. For example, as it can be seen that except for three months –July, August & September, the number of dengue incidences in 2009 have been roughly higher than the year 2010.Such information urges us to check our dataset more thoroughly for these particular three months. It should be noted that only for the demonstration purpose we are considering the monthly dengue cases that have occurred in the two years (2009-2010), just to see if there exits some correlation between these factors. After knowing all vital information about our dataset, we are prepared to carry out the next step of data preparation stage , called data cleaning which is explained in the below section.

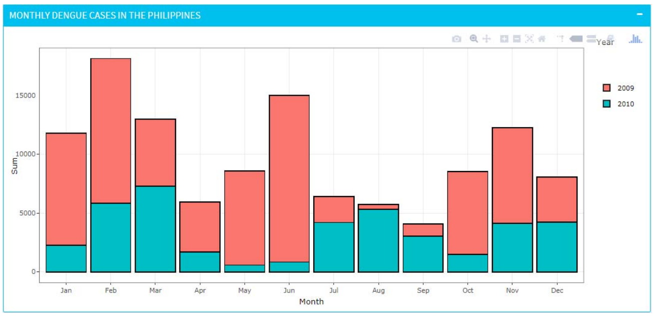


Fig. 1. A bar plot showcasing the monthly dengue cases in Philippines

C. Data Cleaning

Most often after data has been collected, data screening, should take place to make sure that data to be analyzed is as ‘precise’ as it can be. The purpose of this step is to identify incomplete, inaccurate or unrelated data points which are then replaced, altered or deleted. Inaccurate or inconsistent data can cause a number of issues which can lead to drawing of incorrect conclusions. Therefore, data cleaning becomes a chief constituent in data analysis situations.

Keeping this in mind, we performed the following assessments for our analysis:-

1) Missing Value Treatment: Missing data which refers to an empty data value being stored for a variable in an observation is a common phenomenon in real world problems. The presence of missing values in a dataset is likely to affect the data insights as well as the performance of our predictive model, making it a crucial step of data exploration and data preparation stage. Being such an excruciating pain, it becomes important to handle them effectively so as to reduce bias and to produce powerful models. The question now is how to handle any missing data point in our dataset? In general, there are various methods to deal with them such as deletion, imputation, prediction model methods etc. In our dataset also we had missing values for which we used all these methods, for example, missing values in columns like minimum temperature, maximum temperature etc. were replaced by min/max of previous three days value of these columns. Similarly, missing point in snowfall data column was replaced by sum of snowfall in last three days. Likewise, all the missing data points were being taken care of.

2) Second assessment which we performed is *feature engineering,* which is an exercise of extracting extra information from the existing dataset. This step itself is divided into two parts: - Variable transformation & Variable creation. Former is usually done to change the scale, distribution or relationship of a variable while the latter process helps in generating new variables out of the existing variables. So for our case, we generated new variables such as last three days temperature (denoted as L3_temp), last three days rainfall (denoted as L3_rain) etc. The purpose of these newly generated variables is that the dengue infected region is likely to get affected by these variables as well, along with the present day weather elements. Hence this step helped to highlight the hidden relationship of a variable with respect to our target variable.

So these above mentioned data preparation steps helped us to fetch useful information out of our raw dataset. Now using this information, we’ll discuss the application of machine learning algorithms to our dataset in the next subsection.

D. Model Development

With the help of weather API key and pre-processed data obtained from step 4.3, we’ll solve our problem statement using a supervised machine learning algorithm called Random Forest. The following is the description of this algorithm.

Random Forest: a supervised machine learning algorithm

Used for both regression and classification tasks, random forest is an ensemble learning technique that function by creating a multiple of decision trees during training period and outputting the class that is mean (in case of regression) or mode of the classes(for classification) of individual trees. In our experiment, we’ll be using this supervised learning method for regression purpose since we are dealing with a labelled data here. To solve our problem as mentioned in section 3, we first divided our pre-processed dataset into train and test followed by tuning of parameters and finally applying random forest algorithm in R which is a programming software tool for statistical computing and graphics.

V. RESULTS

The application of Random forest technique gave the expected number of dengue occurrences against different weather conditions using which we calculated the prediction score of our benchmark model .This score is then used to calculate the following metrics for our model:-Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. *Variable Importance Plot*

This plot identifies the important features that are closely associated with the target variable and contribute more for variation of the outcome variable. Fig. 2 shows this plot .Going by our dataset, month variable (which is the factor variable) is the topmost contributor which impacts our analysis of dengue prediction. Though this is expected but the important point to highlight is that the topmost two weather variables which impacts the analysis most are pressure and rainfall, followed by other weather factors. It should be noted that these factors doesn't tell whether there would be more or less dengue incidences, but tell which all parameters are important from prediction point of view.

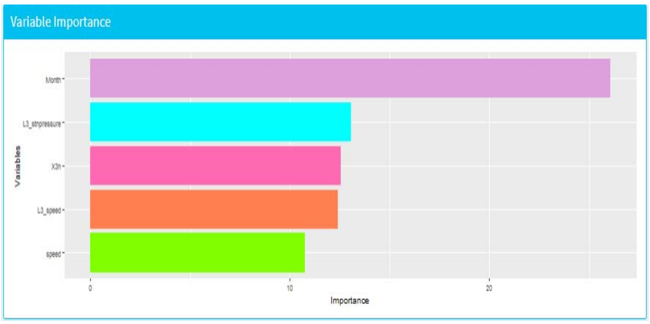


Fig. 2. Variable Importance Plot

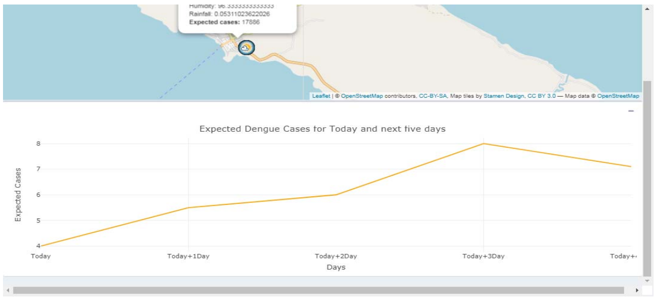
B. *Root Mean Square Error*

This performance metric tells how well our random forest model is able to predict the test set outcomes. Model with smallest value of RMSE is chosen as an optimal model in terms of performance. For our model, this value is close to 1, which is a positive signal towards our model prediction.

With these two metrics, we were able to find the expected number of dengue cases given the weather conditions of Philippines region for the test dataset.

VI. REAL-TIME MODEL

The above section explains our methodology for predicting the dengue cases given the two years weather data for Philippines. We'll apply the same methodology to determine the number of cases for the upcoming five days given five-day weather forecast. Here, the five-day weather forecast refers to the weather conditions that are likely to occur in Philippines for the next five days starting from today. An important thing to note is that this weather data is subject to real-time. So if we'll perform this analysis today, then it would automatically take the weather conditions of the upcoming five days starting from today. Similarly, if we'll run this analysis after one month, then the five-day weather conditions would be for the five days when this model was being run and hence the values would again get change. Thus, using this real-time weather data, along with the model created above, we are forecasting the dengue incidences for the upcoming five days over a google map. Basically, this live model shows two things: 1) Current weather conditions for a city on map and 2) Expected dengue cases for the next five days. The following figure (Fig.3) shows how dashboard for real-time model looks like:


Humidity: 96.3333333333333
Rainfall: 0.05311023622026
Expected cases: 17886

Leaflet | © OpenStreetMap contributors, CC-BY-SA, Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap
Expected Dengue Cases for Today and next five days
Expected Cases
8
7
6
5
4
Today
Today+1Day
Today+2Day
Today+3Day
Today+
Days

Fig. 3. Live Demo model

The above figure shows that today(i.e. the day when we ran this model last), the expected number of dengue cases are four while for the next four days, the range lies between five and eight. And being a real-time model, these figures would keep on changing according to the date on which we would run this model.

VII. CONCLUSION

Under the impetus of global climate variability, there has been drastic increase in the outbreak and spread of various infectious diseases. Meteorological factors like temperature, rainfall, humidity etc., can impact pathogens, hosts and disease vectors, and thus affects their spread patterns severely. Keeping this in mind, we developed an early Warning system (EWS) associating infectious diseases with weather events, preparing society against any health related risks arising due to climate change. This experimentation provides a platform to practitioners, academics and decision makers for sharing and examining how climatic factors is affecting human population and the impact it will have in the future(real-time). In this way, it provides a right set of circumstances to maximize people's resilience and minimize their health risks in real time. Also, though this study revolves around only one disease and one geographic area but this same study can be leveraged to other such vector-borne diseases and locations. Thus, this analysis is applicable globally and enables comprehensive scenarios of daily disease outbreaks to be explored using real-time weather API, allowing disease control measures to be effectively targeted, timed and implemented.

Further steps are required to improvise this analysis:

1. Addition of more explanatory variables such as environmental variables, is required.

2. Also, the dataset on which this analysis has been performed is for two years .So expanding the size of the dataset (for the country-Philippines) would be the next priority.

REFERENCES

[1] Jetten, Theo H., and Dana A. Focks. "Potential changes in the distribution of dengue transmission under climate warming." *The American journal of tropical medicine and hygiene* 57.3 (1997): 285-297.

[2] Reiter, Paul. "Climate change and mosquito-borne disease." *Environmental health perspectives* 109.Suppl 1 (2001): 141.

[3] Thomson, M. C., et al. "Malaria early warnings based on seasonal climate forecasts from multi-model ensembles." *Nature* 439.7076 (2006): 576.

[4] Yang, H. M., et al. "Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue." *Epidemiology & Infection* 137.8 (2009): 1188-1202.

[5] Lambrechts, Louis, et al. "Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti." *Proceedings of the National Academy of Sciences* 108.18 (2011): 7460-7465.

[6] Epstein, Paul R. "Climate and health." *Science* 285.5426 (1999): 347-348.

[7] Rodó, Xavier, et al. "Climate change and infectious diseases: can we meet the needs for better prediction?." *Climatic change* 118.3-4 (2013): 625-640.

[8] Ostfeld, Richard S., and Jesse L. Brunner. "Climate change and Ixodes tick-borne diseases of humans." *Phil. Trans. R. Soc. B* 370.1665 (2015): 20140051.

[9] Enscore, Russell E., et al. "Modeling relationships between climate and the frequency of human plague cases in the southwestern United States, 1960-1997." *The American Journal of Tropical Medicine and Hygiene* 66.2 (2002): 186-196.

[10] Hii, Yien Ling, et al. "Forecast of dengue incidence using temperature and rainfall." *PLoS neglected tropical diseases*6.11 (2012): e1908.

[11] Tong, Shilu, et al. "Climate variability and Ross River virus transmission." *Journal of Epidemiology & Community Health*56.8 (2002): 617-621.

[12] Abeku, T. A., et al. "Effects of meteorological factors on epidemic malaria in Ethiopia: a statistical modelling approach based on theoretical reasoning." *Parasitology* 128.6 (2004): 585-593.

[13] Yang, Guo-Jing, et al. "A growing degree-days based time-series analysis for prediction of Schistosoma japonicum transmission in Jiangsu province, China." *The American journal of tropical medicine and hygiene* 75.3 (2006): 549-555.

[14] Ngeleja, Rigobert C., Livingstone S. Luboobi, and Yaw Nkansah-Gyekye. "The effect of seasonal weather variation on the dynamics of the plague disease." *International Journal of Mathematics and Mathematical Sciences* 2017 (2017).

[15] Iacono, Giovanni Lo, et al. "Challenges in developing methods for quantifying the effects of weather and climate on water-associated diseases: A systematic review." *PLoS neglected tropical diseases* 11.6 (2017): e0005659.

[16] Wu, Yan, et al. "Mining weather information in dengue outbreak: predicting future cases based on wavelet, SVM and GA." *Advances in Electrical Engineering and Computational Science*. Springer, Dordrecht, 2009. 483-494.

[17] Hales, Simon, et al. "El Niño and the dynamics of vectorborne disease transmission." *Environmental Health Perspectives*107.2 (1999): 99.

[18] Gagnon, Alexandre S., Andrew BG Bush, and Karen E. Smoyer-Tomic. "Dengue epidemics and the El Niño southern oscillation." *Climate Research* 19.1 (2001): 35-43.

[19] Cazelles, Bernard, et al. "Nonstationary influence of El Nino on the synchronous dengue epidemics in Thailand." *PLoS medicine* 2.4 (2005): e106.