# Dengue Outbreak Prediction Using Data Mining Techniques

Asha Bharambe and Dhananjay Kalbande

April 20, 2020

# Dengue Outbreak Prediction Using Data Mining Techniques

Asha Bharambe[1], Dr. Dhananjay Kalbande[2]
[1] Department of Information Technology, V.E.S.I.T, Mumbai, India
[2] Department of Computer Engineering, S.P.I.T, Mumbai, India

**Abstract.** The incidence of dengue has grown dramatically around the world in the past decade. Preventive measures should be adopted to reduce the number of incidences and deaths caused by Dengue. Measures for early case detection, improved outbreak detection and prevention techniques are required to be implemented. We have implemented several data mining techniques for prediction of dengue outbreaks.

**Keywords:** Data mining, Time Series, Dengue, Prediction.

## 1. Introduction

Outbreaks have a massive burden on public health systems, populations, and economies in most countries of the world. In the recent years, dengue has grown to be major epidemic across the world.

According to World Health Organization (WHO), the incidence of dengue has grown dramatically around the world in the recent decade. The disease is now endemic in more than 100 countries [1]. It is estimated that there are over 50-100 million cases of dengue worldwide each year, with 3 billion people living in dengue endemic countries. The number of cases reported increased from 2.2 million in 2010 to 3.2 million in 2015. An estimated 500 000 people with severe dengue require hospitalization each year, a large proportion of whom are children. About 2.5% of those affected die. Not only is the number of cases increasing as the disease spreads to new areas, but explosive outbreaks are occurring. Figure 1 shows a plot of the dengue alerts for the past one week.

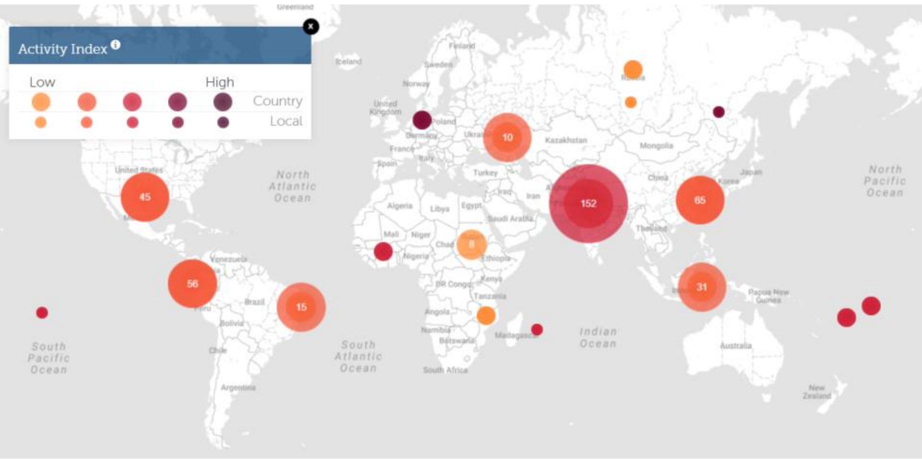

Figure 1: Dengue alerts in past one week

Source: www.healthmap.org/dengue/en/

Given the high morbidity and mortality of dengue, it is of importance that preventive measures should be adopted to reduce the incidences and deaths caused by Dengue. Measures for early case detection, improved outbreak detection and prevention techniques are required to be implemented.

## 1.1. Epidemiology

Dengue viruses are the causative agents of dengue fever (DF) and dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS) in humans. Arboviral infections are usually sensitive to changes in rainfall and temperature.

The transmission of dengue virus depends upon biotic and abiotic factors. Biotic factors include the virus, the vector and the host, whereas abiotic factors include temperature, humidity and rainfall.

Dengue is a vector borne viral infection which spreads due to Aedes Aegypti mosquito. There are four virus serotypes, which are designated as DENV-1, DENV-2, DENV-3 and DENV-4. Infection with any one serotype confers lifelong immunity to that virus serotype. Although all four serotypes are antigenically similar, they are different enough to elicit cross-protection for only a few months after infection by any one of them. Secondary infection with another serotype or multiple infections with different serotypes leads to severe form of dengue (DHF/DSS). Dengue viruses of all four serotypes have been associated with epidemics of dengue fever (with or without DHF) with a varying degree of severity.

Dengue is prevalent throughout the tropics, with risk factors influenced by local spatial disparities of rainfall, temperature, relative humidity, degree of urbanization and quality of vector control services carried out in the areas [1].

Currently, no licensed vaccine is available against dengue infection. Hence, early prediction and prevention is a viable solution. This disease can be prevented by having an active surveillance system that can identify the areas of infection and predict the future infections.

Early warnings for outbreaks can be used to inform stakeholders and help them plan properly to control the disease. They can be helpful to provide insights into regional and national public health burden imposed by dengue.

The prediction can be achieved through data mining and statistical techniques.

### 1.2 Disease Surveillance

Surveillance helps in monitoring the dengue situation in the area and is carried out on regular intervals.

The Epidemiological surveillance enables the collection of the number of cases and deaths in the given area due to dengue. It helps in detection of epidemics, to measure the disease burden and monitor trends in the propagation of dengue over time.

Entomological surveillance helps in the identifying the vector density in an area by performing a larval surveillance during the pre-monsoon and monsoon period. It helps to find out the extent of occurrence of the vectors in selected high-risk area. Measures like House Index, Breteau Index and Container Index can be used to monitor the vector population. [3]

## 2. Related work in outbreak prediction

### Data Mining and Statistical Techniques

The data mining field has emerged as a combination of multiple areas like artificial intelligence, machine learning, databases, data visualization, statistics etc. Data mining techniques helps in analyzing the data and identifying unknown relationships amongst the data which helps in decision making. It also helps in identifying the repetitive patterns in the data and aids in predicting the future outcomes based on the patterns and relationships. These techniques have been used in the outbreak prediction in past. Most commonly used techniques are classification and regression, time series analysis and hybrid models.

Knowledge of the dynamic structure of underlying data collected at regular time intervals can be useful in generating forecasts through data modelling. This can be achieved either by self-projecting or cause-effect modelling techniques. Self-projecting approaches produce predictions using only time series of the activity to be forecasted while the latter relies on relationships between the time series to be forecasted and one or more series that influence it.

Naïve method predictions are based on the last observed value in the series and give no importance to the pervious observations.

Decomposing time series models are those which separates time series into three components: trends, seasonal, and random. Some of the classical methods for time series decomposition are autocorrelation regression (AR), Moving Averages(MA) models, combination of these (ARMA, ARIMA), Seasonal & Trend Decomposition using Loess (STL) method and exponential smoothing methods like Holt's method.

Holt’s method takes weighted averages of past observations of data with trend component, with the weights reducing exponentially as the observations get older i.e. the most recent observations are associated with higher weights.

Various techniques have been used to predict the outbreaks for diseases[4]. Regression models[5], time series models[6] haven been used along with the metrological data to predict the outbreak in various parts across the world.

## 3. Methodology

### 3.1 Strategy for dengue prevention.

Dengue prevention can be done with the help of controlling the spread of the disease. Thus for early detection the following strategy is suggested.

1. Apply clustering techniques on the surveillance data to identify the high-risk areas.
2. Predict the morbidity rate in the area.
3. Enable early detection of the cases.

### 3.2 Data collection

For the implementation purpose, we have used the dataset available from the source http://dengueforecasting.noaa.gov/. The dataset defines the dengue dataset at the geographical location Iquitos, Peru, USA and the climatic factors like the rainfall, temperature, humidity, population data. The occurrence of dengue over the period of 10 years is shown in Figure 2

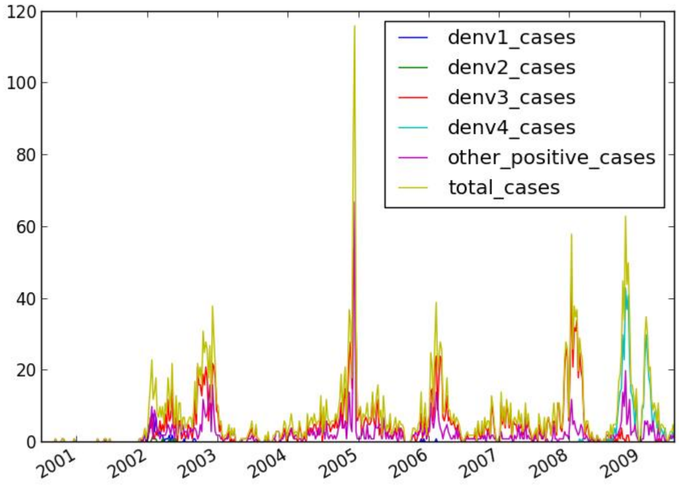

**Figure 2: Plot of dengue cases from year 2000 to year 2009**

### 3.3 Methods and Discussion

R language is used to build the prediction model. A preliminary analysis was first conducted on the dataset to describe and investigate the nature of the trend characterizing the number of cases in a region. The data is represented as a time series data for total occurrences of all dengue cases. Figure 3 shows the decomposition of the time series into following seasonal, trend and random plots.

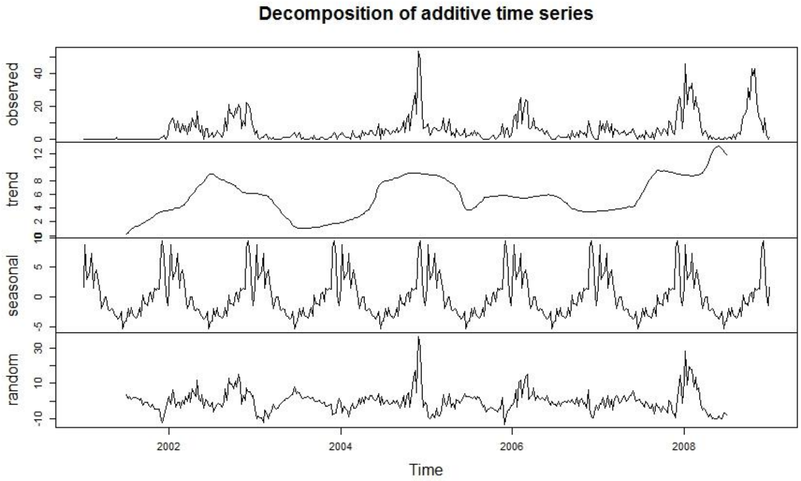

**Figure 3: Decomposition of the time series**

In order to determine the trend, linear, quadratic, log-linear and log-quadratic regression models were fitted and compared. A seasonal autoregressive integrated moving average (SARIMA) model (0,1,1)(2,2,1) was fitted with the data from 2000 to 2008 and tested on data from 2000 to 2012. The plot of the predicted vs actual values is shown in figure 4.

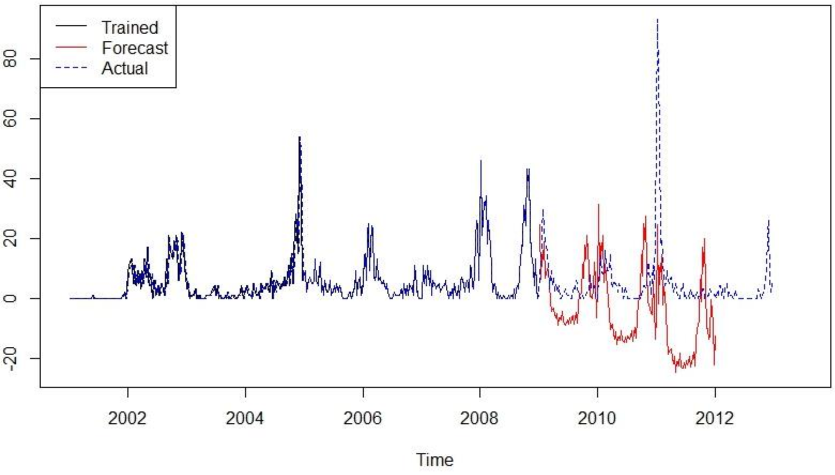

**Figure 4: Forecasted vs Actual data of SARIMA (0,1,1)(2,2,1) model**

STL models were implemented and their plots is shown in figure 5

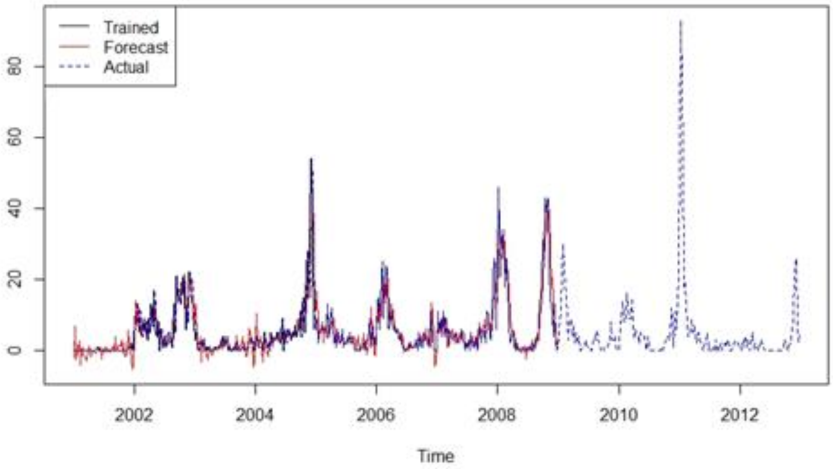

**Figure 5: Forecasted vs Actual data of STL model**

## 4. Results

The predicted values using various algorithms are compared based on following performance measures

- Mean absolute error(MAE),
- Root Mean Square Error (RMSE)

Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for continuous variables.

**Mean Absolute Error (MAE):** MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \overline{y}_i |$$

**Root mean squared error (RMSE)**: RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

The methods were implemented and the performance was calculated as shown in Table 1. The STL method provided with the best results for the prediction.

| **Model** | **Data** | **RMSE** | **MAE** |
|---|---|---:|---:|
| **SNaive** | Test | **11.92** | **7.80** |
| **STL- Seasonal decomposition** | Train | 4.57 | 2.98 |
| | Test | **6.72** | **4.81** |
| **HoltWinters** | Test | **7.94** | **7.00** |
| **SARIMA (0,1,1) (2,2,1)** | Test | **9.28** | **8.22** |
| **SARIMAX (0,1,1) (2,2,1)** | Test | 16.03 | 12.24 |

Table 1: Performance of prediction techniques

## 5. Conclusion

This research proposes to help healthcare providers by providing a framework which would provide them with the surveillance data. It will also offer a prediction model which will empower them with improved decision making in taking preventive and control measures.

## 6. References

1. World Health Organization, http://www.who.int/denguecontrol/en/
2. National Vector Borne Disease Control Programme, http://www.nvbdcp.gov.in/den-cd.html
3. Martinez R. Working paper 7.2. Geographic information system for dengue prevention and control. In: WHO/TDR. Report of the Scientific Working Group meeting on Dengue, Geneva, 1-5 October 2006. Geneva, 2007. Document no. TDR/SWG/07. pp. 134–139
4. Bharambe, Asha A. and Dhananjay R. Kalbande. “Techniques and Approaches for Disease Outbreak Prediction: A Survey.” *WIR '16* (2016).
5. Khormi HM, Kumar L, Elzahrany RA (2013) Regression Model for Predicting Adult Female Aedes aegypti Based on Meteorological Variables: A Case Study of Jeddah, Saudi Arabia. J Earth Sci Clim Change 5: 168. doi:10.4172/2157-7617.1000168
6. Fabgge Li, Peixian Luan, “ARMA Model for Predicting the Number of New Outbreaks of Newcastle Disease during the Month”, Computer Science and

Automation Engineering (CSAE), 2011 IEEE International Conference on (Volume:4 ) IEEE 2011, ISBN - 978-1-4244-8727-1, June 2011