International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 9443-9455
© Research India Publications
http://www.ripublication.com

Early Warning System For Dengue Outbreak – A Preliminary Approach Using Time Series Forecasting

Dr.L.S Jayashree[1] , Lakshmi Devi.R[2], Dinesh.R[3]
[1]Dean – Computer Science & Engineering,
R.V.S College of Engineering, Coimbatore, Tamil Nadu, India
Email Id: jayashreecls@yahoo.co.in
[2]Asst.Professor, Dept. of Computer Applications,
S.A Engineering College, Chennai, Tamil Nadu, India
Email Id: lakshmidevi2305@gmail.com
[3]Dept. of Computer Science & Engineering,
R.V.S College of Engineering, Coimbatore, Tamil Nadu, India

Abstract

Dengue is one of the challenges faced by tropical countries such as India. The impacts of climate changes can affect dengue outbreak. From the previous studies it is examined that there is an association between metrological variables and dengue incidence using Time series analyses. The proposed study is to explore a systematic approach that provides an early warning system for dengue outbreak of a given region. . The time series data is decomposed and estimated the trend, seasonal and irregular compounds. Time series analysis using ARIMA and SARIMA model along with temperature variants is found to be effective for dengue predication. The prediction is based on the benchmark data of Dengue incidence and metrological data using R-tool version 3.0.2. Experimental result shows that the metrological variables (Maximum temperature, Humidity and Rainfall) significantly influence the dengue incidence for the given dataset. Error values of the SARIMA model provides comparatively lower with respect to ARIMA.

Keyword: Time Series analysis, ARIMA, SARIMA model, Dengue prediction, Regression.

Introduction

Dengue infection is caused by four antigenically distinct serotypes of the dengue virus (DENV1, DENV2, DENV3, and DENV4), and its main vector is the *Aedes aegypti* mosquito. Dengue is considered the most significant arbovirus that affects humans. Dengue is estimated to annually cause 390 million infections, including 96 million cases of classical dengue and 20,000 deaths caused by dengue [2].

In many parts of the tropics and subtropics, dengue is endemic, that is, it occurs every year, usually during a season when *Aedes* mosquito populations are high, often when rainfall is optimal for breeding. These areas are, however, additionally at periodic risk for epidemic dengue, when large numbers of people become infected during a short period [1].

Various analytic techniques like Time series analyses are often used on metrological and Dengue incident data to provide insights on different patterns within them which may be useful for purposes such as prevention and mitigation. Even though daily outcome data are anticipated for time series analysis, obtaining such data from most of the developing countries is impossible. Hence, most time series analyses use monthly or annual data.

Tools and Techniques

The techniques used so far for dengue prediction and forecasting involved statistical, data mining and machine learning methods. The methods such as Decision Trees, Support Vector Machine, Genetic Algorithms, Fuzzy sets, Neural Networks and Rough sets are classification methods that are found to be effective in diagnosis and prognosis of the Arbovirus-Dengue. Though these methods can we used to predict the dengue occurrence, some of them do not capture the seasonality pattern that is found to involve in it. The seasonal variation of the ARIMA model is used with effectiveness for Dengue outbreak prediction along with the seasonal variants.

The language R was used exhaustively used for the basis Statistical and time series analysis. R is a multi-paradigm language which includes array, object-oriented, imperative, functional, procedure and reflective paradigms. R is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls and surveys of data miners are showing R’s popularity has increased substantially in recent years.

Statistical Model For Dengue Forcasting

(i) ARIMA Model

The ARIMA (Auto-Regressive Integrated Moving Average) model is the most general class of model for forecasting a time series which can be stationarised by transformation such as differencing and logging. This model is mainly used for non-stationary time series data.[4]

A non-seasonal ARIMA model is classified as an ARIMA (p,d,q) model, where
p is the number of autoregressive terms
d is the number of non-seasonal differences
q is the number of lagged forecast errors in the prediction

(ii) SEASONAL ARIMA Model

The seasonal part of an ARIMA model has the same structure as the non-seasonal part. A seasonal ARIMA model is classified as an ARIMA (p,d,q) x(P,D,Q).

Where
P=number of seasonal autoregressive (SAR)
D=number of seasonal differences
Q=number of seasonal moving average (SMA)

(iii) Plotting Time Series

In R tool ts.plot () method is used to view the time series. This would allow seeing the overall trend and nature of the time series. The time series plot of benchmark data of dengue incidence from 2010 to 2013 is given below

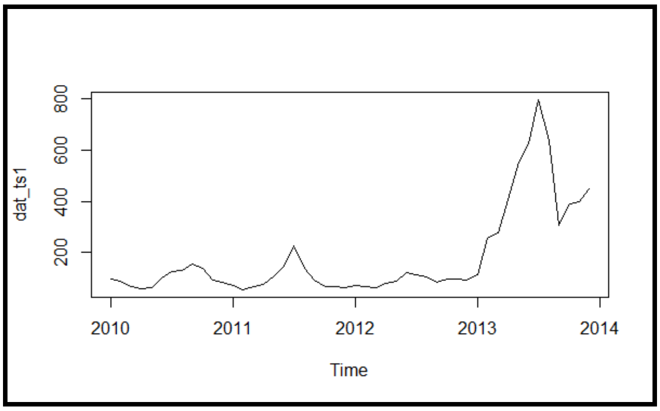

Figure 1: Time Series of Dengue incidence from 2010 to 2013

There seems to be seasonal variation in the number of dengue incidence every month and there is a peak every summer. The time series could be described using an additive model, as the seasonal fluctuations are roughly constant in size over time and do not seem to depend on the level of the time series, and the random fluctuations also roughly constant in size over time. The transformation is done using the calculated natural log of the original data and obtains the plot as below.

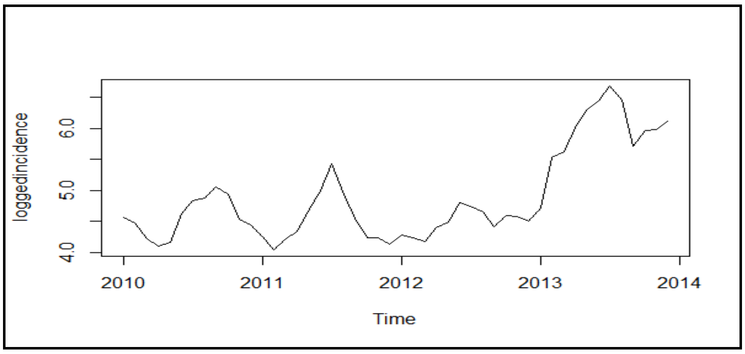

Figure 2: Logged Time Series

The size of the seasonal fluctuations and random fluctuations in the log-transformed time series seem to be roughly constant over time, and do not depend on the level of the time series. The log-transformed time series can be described using an additive model.

(iv)Decomposing Time Series

Decomposing a time series is separating into its constituent components, which are usually a trend component and an irregular component, and if it is a seasonal time series, a seasonal component.

A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components: that is, estimating these three components.

To estimate the trend component and seasonal component of a seasonal time series that can be described using an additive model, use the “decompose ()” function in R. This function estimates the trend, seasonal, and irregular components of a time series that can be described using an additive model. The function “decompose ()” returns a list object as its result, where the estimates of the seasonal component, trend component and irregular component.

The time series of the number of dengue incidence per month is seasonal with a peak every summer and described using an additive model since the seasonal and random fluctuations seem to be roughly constant in size over time.

Table 1: Seasonal Components

Year	Jan	Feb	Mar	Apr	May	Jun
2010	-68.250579	-42.6728	-41.6499	3.81956	58.0897	97.10151
2011	-68.250579	-42.6728	-41.6499	3.81956	58.0897	97.10151
2012	-68.250579	-42.6728	-41.6499	3.81956	58.0897	97.10151
2013	-68.250579	-42.6728	-41.6499	3.81956	58.0897	97.10151
	Jul	Aug	Sep	Oct	Nov	Dec
2010	60.545255	27.38206	7.945949	-8.7728	-36.0353	-57.5027
2011	60.545255	27.38206	7.945949	-8.7728	-36.0353	-57.5027
2012	60.545255	27.38206	7.945949	-8.7728	-36.0353	-57.5027
2013	60.545255	27.38206	7.945949	-8.7728	-36.0353	-57.5027

The estimated seasonal factors are obtained for the months January-December, and are the same for each year. The largest seasonal factor is for June about 97, and the lowest is for January about -68, indicating that there seems to be a peak in the dengue incidence in June and a trough in dengue incidence in January each year. Using the “plot ()” function estimated trend, seasonal, and irregular components of the time series are shown below.

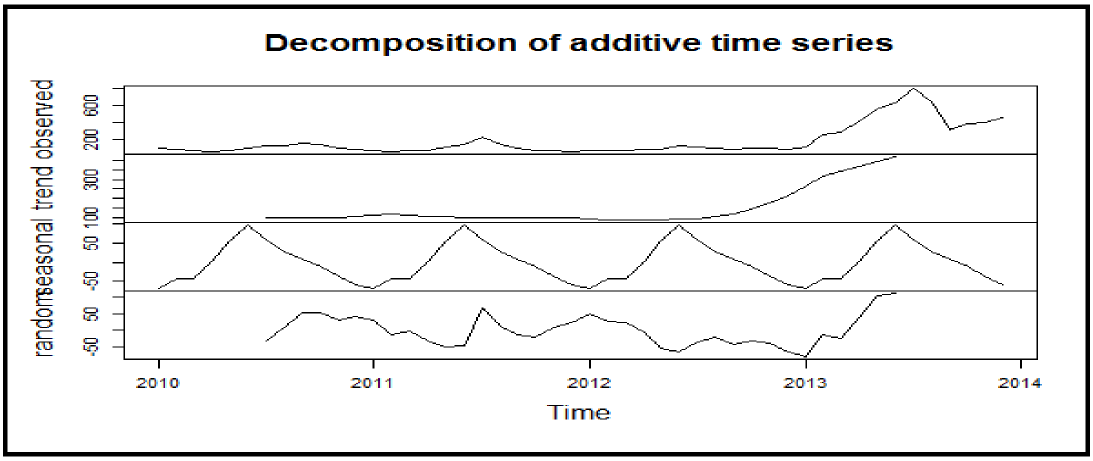

Figure 3: Decomposition of Time Series

The plot above shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom). The estimated trend component shows a slight increase and decrease from 2010 to 2012 after which there is a steady increase from then on.

(v) Seasonally Adjusting

A seasonal time series can be described using an additive model, seasonally adjust the time series by estimating the seasonal component, and subtracting the estimated seasonal component from the original time series. The plot obtained after the adjustment is givenbelow.

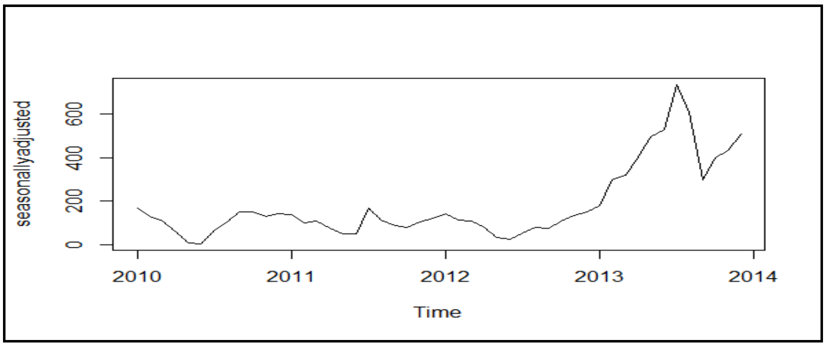

Figure 4: Seasonally Adjusted Time Series

The seasonal variation has been removed from the seasonally adjusted time series. The seasonally adjusted time series now just contains the trend component and an irregular component.

Regression Model

Regression analysis is a statistical process for estimating the relationships among variables. There are many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.[5]

(i) Univariate Regression

In univariate regression the variation of a dependent variable with respect to one independent variable is explained. This established, a relationship between the dependent variable with independent variable as an equation of a straight line

$Y=a+bX$………………….. (1)

Where,

a is the intercept value

b is the slope value

Y is the dependent variable

X is the independent variable

(ii) Multivariate Regression

The multivariate regression explains the variation of a dependent variable with respect to more than one independent variable. The number of the independent variables and type are determined by domain study and theory of the problem.

$Y=a+b_0X_1+ b_1X_2+\ldots\ldots b_{n-1}X_n\ldots$ (2)

Where,

a is the intercept value

b_0 is the coefficient value corresponding to X_1

b_1 is the coefficient value corresponding to X_2 and so on…

The regression model used along with the variables is evaluated using the summary function. The summary function with regression displays the information: *Residuals, Significance Stars, Estimated Coefficients, Standard Error of the Coefficient Estimate, t value of Coefficient Estimate, Variable p-value, Significance Legend, Residual std Error/Degrees of Freedom, R squared, F statistic & Resulting p value*

The above informations analyse how well the one or more independent variables make changes on dependent variable.

(i) Residuals

The residuals are the difference between the actual value and predicted value from regression.

Table 2: Residual – Univariate Regression

Independent Variable	**MIN**	**MEDIAN**	**MAX**
Mean Temperature	-140.56	-79.56	608.99
Max Temperature	-162.48	-77.08	593.15
Min Temperature	-194.28	-61.47	564.56
Humidity	-243.49	-35.21	533.27
Rainfall	-151.29	-71.5	610.89

Table 3: Residual – Multivariate Regression

Independent Variables	**MIN**	**MEDIAN**	**MAX**
Mean Temperature+ Max Temperature	-224.58	-56.49	487.49
Mean Temperature+ Max Temperature+ Min Temperature	-260.17	-43.62	440.39
Mean Temperature+ Max Temperature+ Min Temperature+ Humidity	-263.9	-22.13	393.74
Mean Temperature+ Max Temperature+ Min Temperature+ Humidity+ Rainfall	-254.8	-27.48	343.25
Max Temperature+ Humidity+ Rainfall	-219	-21.26	385.95

The result shows that the residual of the univariate regression with respect to the individual variants quite high.

(ii) Significance Stars

The significance stars are shorthand for significance levels, with the number of asterisks displayed according to the p-value computed. Higher the significance starts higher it is unlikely that no relationship exists between the dependent and the independent variables. The significance level here indicates how much related the variables are in the regression model.

Table 4: Significance stars - Univariate Regression

Independent Variable	**Intercept**	β_0
Mean Temperature	-	-
Max Temperature	-	-
Min Temperature	-	-
Humidity	**	**
Rainfall	*	-

Table 5: Significance stars - Multivariate Regression

Independent Variables	**Intercept**	β_0	β_1	β_2	β_3	β_4
Mean Temperature+ Max Temperature	-	*	**			
Mean Temperature+ Max Temperature+ Min Temperature	-	-	*	-		
Mean Temperature+ Max Temperature+ Min Temperature+ Humidity	*	-	*	-	**	
Mean Temperature+ Max Temperature+ Min Temperature+ Humidity+ Rainfall	**	-	*	-	***	*
Max Temperature+ Humidity+ Rainfall	***	**	***	*		

The univariate linear regression on the given data set find that the there is some significance only in the case of trying to explain the incidence on the basis of humidity. Thus move on to apply the multivariate regression on the dataset. Here different combination of the independent variables to explain the dengue incidence and it is evident from the above table that the best combination with higher level of significance level for the explanation of the dengue incidence depending upon the data set used is *"Max Temperature + Humidity + Rainfall"*.

Hence proceeded to analyze more on the regression based on the independent variables namely the maximum temperature, humidity and the rainfall of the region.

(iii) Estimated Coefficients

The estimated coefficient is the value of the slope calculated by the regression. The independent variable and their corresponding estimated coefficients are given below.

Table 6: Estimated coefficients - Multivariate Regression

Independent Variables	**Estimated Coefficients**
Intercept	7081.537
Max Temperature	-92.017
Humidity	-51.700
Rainfall	13.980

These estimated values are the approximate of how much change each of the variables considered produce with respect to the dengue incidence. These values can be used to predict or estimate the next value or the number of dengue incidence for the next time period by calculating it with respect to the regression line equation.

(iv)Residual Vs Fitted Plot

The residual Vs fitted plot is used to visualize the regression. It is a scatter plot of residuals on the y axis and fitted values on the x axis. This plot is used to detect non-linearity, unequal error variances, and outliers.

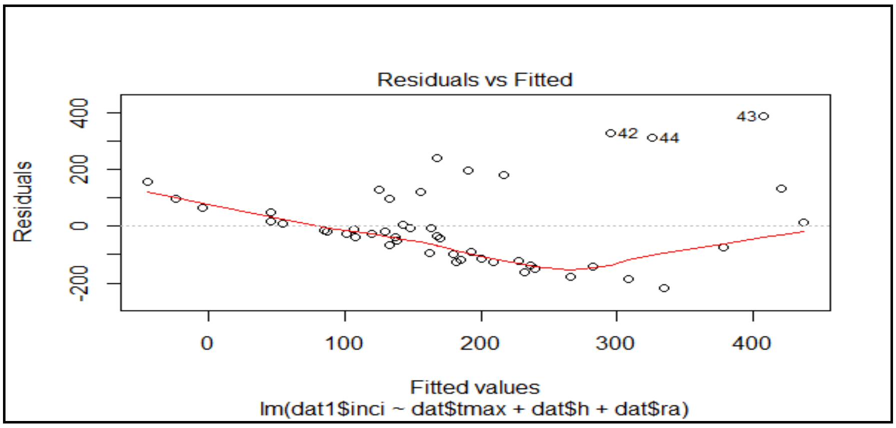

Figure 5: Residual vs Fitted

(v) Normal Q-Q Plot

The normal Q-Q plot is used to check if the residuals are normal. A normal residual indicates than the model suitably recognizes all the dependencies.

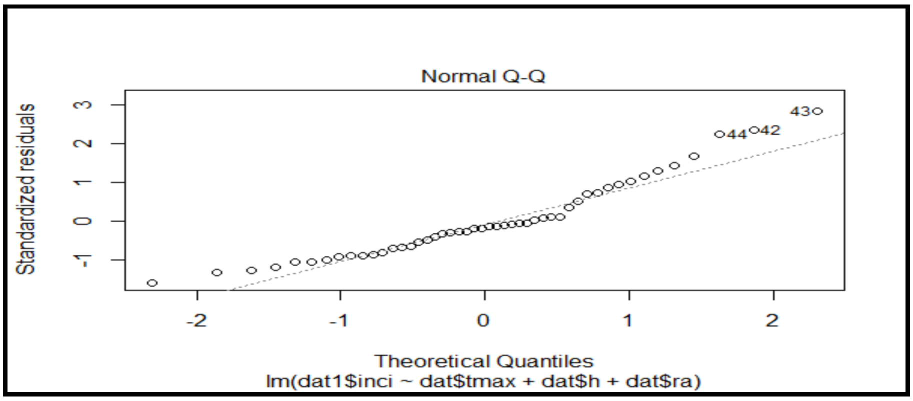

Figure 6: Normal Q-Q

Obtained the plot of the quartiles Vs standardized residuals and it can be seen that the residuals almost plot to a normal curve.

(v) Scale-Location Plot

A scale-location plot is similar to the residuals versus fitted values plot, but it uses the square root of the standardized residuals.

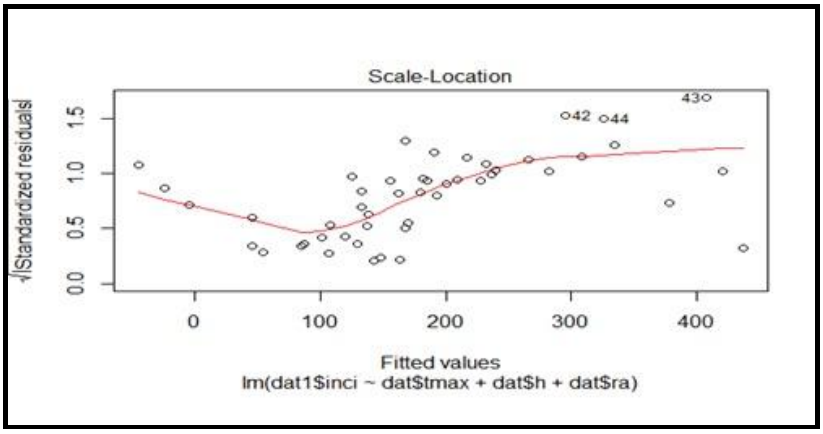

Figure 7: Scale-Location

Thus, analysed the dengue incidence based upon the climatic variables. Selected the suitable variables appropriate for regression and have modelled the multivariate regression based on them.

Results and Discussion

Auto Regressive Integrated Moving Average (ARIMA) models include an explicit statistical model for the irregular component of a time series that allows for non-zero autocorrelations in the irregular component. The auto.arima() function can be used to find the appropriate ARIMA model. The resultant plots and errors are given below

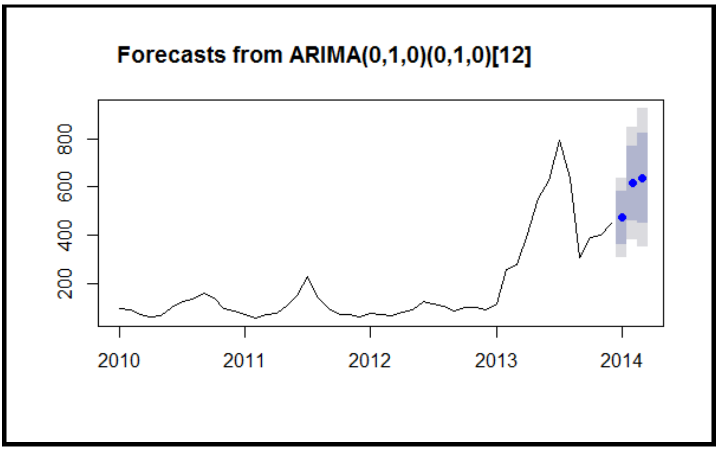

Figure 8: Forecasts from ARIMA

Training set	**Error measures**
ME	8.015842
RMSE	72.44531
MAE	40.1471

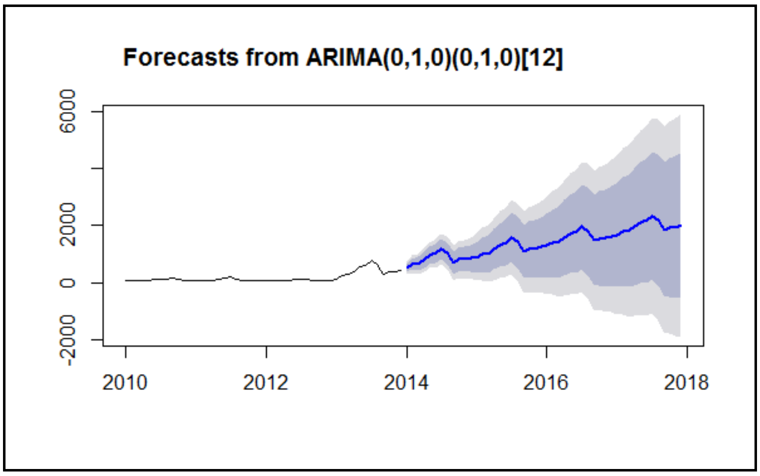

Figure 9: Forecasts from SARIMA

Training set	**Error measures**
ME	8.079875
RMSE	70.87701
MAE	39.90616

The MAE measures for the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables.

The MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

The MAE and RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMS=MAE, then all the errors are of the same magnitude.

Both the MAE and RMSE can range from 0 to infinity. They are negatively-oriented scores: Lower values are better. The SARIMA model provides comparatively lower errors with respect to ARIMA.

Conclusion

The Benchmark data for Dengue records from the year 2010 to 2013 was collected. The error measures for the models used are determined and the predictions are made. The regression models were used to determine the amount of influence each of the temperature variable have on the dengue incidence. The factors were determined to be maximum temperature, humidity and rainfall.

The time series analysis was done on the obtained data set to determine the trend, seasonal and the random components. The decomposition estimates the seasonal factors and obtained for the months January-December. It is observed that the largest seasonal factor is June about 97, and the lowest is for January about -68, indicating that there seems to be a peak in the dengue incidence in June and a trough in dengue incidence in January each year.

When comparing the ME, RMSE, MAE error values the SARIMA model provides comparatively lower errors with respect to ARIMA.

Future Work

The work can be enhanced by the inclusion of other factors that might influence the disease incidence. The methods and the procedure used can also be applied to determine the outbreaks of seasonally occurring diseases. Also it can be used for other non-stationary time series in other fields such the stock prices.

References

[1] WHO – World Health Organization, 2010, “Dengue and dengue haemorrhagic fever [factsheet]”,Geneve.

[2] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Browntein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GRW, Simmons CP, Scott TW, Farrar JJ, Hay SI,2013, “The global distribution and burden of dengue”, 496: 504-7.

[3] Halmar Halide, Rais and Peter Ridd,2011,”Early Warning System for Dengue Hemorrhagic Fever(DHF) Epidemics in Makassar”,Jurnal Matematika Dan Sains, Vol. 16 Nomor 2.

[4] Myriam Gharbi,Philippe Quenel,Joel Gustave,Sylvie Cassadou,Guy La Ruche, Laurent Girdary and Laurence Marrama , 2011, “Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors”, Gharbi et al. BMC Infectious Diseases,11:166.

[5] Kensuke Goto,Balachandran Kumarendran, Sachith Mettanandam Deepa Gunasekara, Yoshito Fujii.,2013,”Analysis of Effects of Metrological Factors on Dengue incidence in Sri Lanka using Time series data”, Plos one, Volume 8, Issue 5.