Time Series Predictive Model for Dengue Cases in Sri Lanka

Dengue disease has been identified as a rapidly developing pandemic-prone viral disease in many countries in the world. In recent decades, dengue incidence has dramatically spread in worldwide. Severe dengue is a leading of serious illness and presently, it affects Asian and Latin American countries. Since particular treatment has not been identified for dengue or severe dengue, early detection is very important to reduce the fatality rates by accessing proper medical care. This study is mainly focus on developing time series predictive model to forecast dengue incidence in future months in Sri Lanka. Mainly, monthly dengue cases reported in Sri Lanka during January 2010 to September 2017 has been used for developing time series model. Here, SARIMA (2, 1, 2) (0, 0, 1)4 model has been selected as the most appropriate model based on the AIC value. Also, the model can be used for predicting number of dengue cases in Sri Lanka if the observations of time series do not indicate unusual dengue incidence in future months. The validation criterion for the fitted model has been satisfied and accuracy of the fitted model was checked using measurement of errors (i.e. RMSE, MAPE etc.) which have indicated as satisfactorily small.

October 2017

1. Introduction

Dengue is identified as a mosquito born viral disease in the world. It is widely spread all over the world. Dengue fever is endemic disease in Sri Lanka and it occurs each year. It is discovered that rainfall mostly influences for spreading dengue mosquitoes. In addition, improper human behavior such as unclean water garbage, standing water pools and other potential places have significantly influenced. Presently it has mostly affected to children and adults in Sri Lanka.

As shown in the figure 1.1 graph, it has depicted 157,997 number of dengue cases reported in the year 2017 and this figure is significantly high compared to the number of dengue cases reported in the previous years. Also when considering the reported dengue cases in 2017, it is approximately three times the dengue cases reported (55,150) in 2016. This considerable change in the year 2017 is mainly caused due to a substantial increase in number of dengue cases from month of January 2017 in Sri Lanka (Figure 1.3 graph).

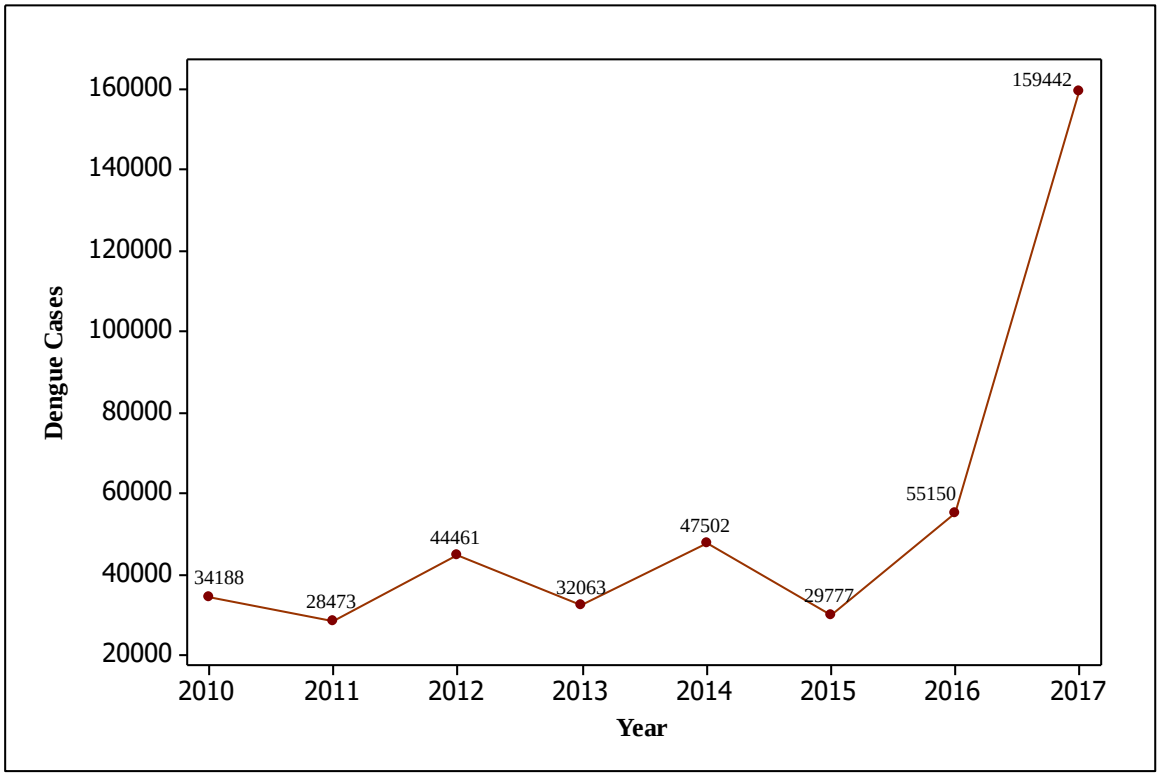

Figure 1.1: Dengue Cases in Sri Lanka during 2010 to 2017

Data Source: Epidemiology Unit in Sri Lanka

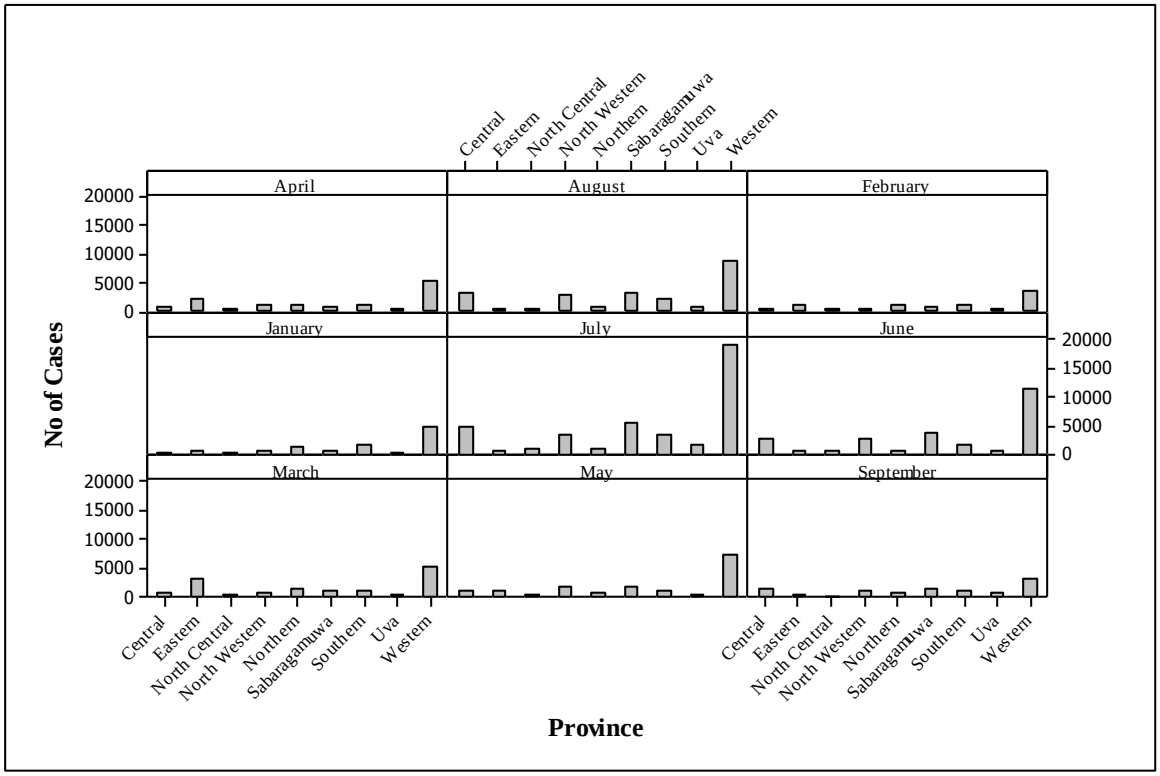

Figure 1.2: Bar Charts of Monthly Dengue Cases in Province wise in Sri Lanka in the year 2017

Data Source: Epidemiology Unit in Sri Lanka

When considering this substantial change during the year 2017, it may occur due to some reasons emerged from the environment. Mainly, lack of cleaning the environment in the area of Colombo may cause to this type of change during the recent time periods. Also, Figure 1.2 graphs clearly depict western province is significantly affected from dengue disease during 2017 year. Apart from rain fall, mostly, improper human habitation towards the environment may influence considerably spread the dengue mosquitoes. Hence, it is important to identify unwanted places where dengue virus is mostly infected and to take necessary actions to solve this problem.

1.1 Objective of the study

Main focus of this study is to develop a predictive model by using monthly dengue cases in Sri Lanka during 2010-2017.

1.2 Significance of the Study

This study is timely appropriate since dengue disease is widely spreading in Sri Lanka and it mainly affects for human beings. Through this study, our aim is to assist Ministry of Health in Sri Lanka in order to implement necessary programs for reducing dengue incidence in future.

2. Methods and Materials

2.1 Data used for the study

Data for this study have been taken from Epidemiology unit, Ministry of Health, Sri Lanka. Monthly dengue cases reported from January 2010 to September 2017 have been used for the model development. In the study, natural log transformation has been taken for the monthly dengue cases with adjusted outliers. Generally, transformation is carried out when time series showing variation increase with the level of the series. Mainly, Transformation is useful to stabilize the variance. Also, after transformation is carried out, transformations are reversed by taking exponential of the forecasts in order to obtain the forecasts on original data. Initially, sample consists of 93 log transformed monthly dengue cases reported from January 2010 to September 2017 with 9 adjusted outliers. It is separated into two data sets as train data and verification data. The train data set have 62 log transformed monthly dengue cases from January 2010 to February 2015 while the verification data set consists of 31 log transformed monthly dengue cases from March 2015 to September 2017.

2.2 Methodology used for Preliminary Analysis and Advanced Analysis

Mainly, Time Series plot and Box plot are used to identify nature of data and to find out outliers in the time series data. Also, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are used for pattern recognition of the time series data.

Autocorrelation

Correlation can be defined as a measure of linear association between two random variables. In time series analysis, the concept of correlation is very important. If X_t (t=1,2,3,4,……) is taken as a sequence of random variables, any pair of these random variables may be correlated. Autocorrelations is referred to as a correlation between X_t and X_{t-j}. It means correlation between the time series and its own past.

Suppose $Corr\ (X_t, X_{t-1}) = \rho_1$. Since ρ_1 represents a correlation, ρ_1 should be between -1 and +1. Here, the larger ρ_1 explains the stronger linear association between X_t and X_{t-1}.

White Noise

A stationary time series ε_t is said to be white noise, if $Corr\ (\varepsilon_t, \varepsilon_s) = 0$ for all t≠ s. Hence, ε_t is a sequence of uncorrelated random variables with constant variance and constant mean.
White Noise Series is denoted as follows.
$\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$

Backward shift Operator

The backward shift operator is a useful operator in time series analysis. It is defined as follows.
$B^k X_t = X_{t-k}$ where k = 1,2,3,4,…….

Here, B operating on X_t which has effect on shifting the data back k periods. Also, terms with backward shift operators can be multiplied together.

Difference Operator

The difference operator is denoted as follows.
$\nabla = 1 - B$

2.2.1 Theory behind Seasonal Auto Regressive Integrated Moving Average Model

It has been carried out Multiplicative Seasonal Auto Regressive Integrated Moving Average (SARIMA) Model for the log transformed data. Following has mentioned the theory behind SARIMA Model according to Shumway and Stoffer (2017).

Seasonal ARIMA is a modification made to the ARIMA model in order to account for seasonal and non-stationary behavior of the time series. When seasonal pattern is available in a time series, it can be generalized the ARMA model for stationary series with both the regular dependence and seasonal dependence. Here, regular dependence is involved with the measurement intervals of the time series whereas seasonal dependence is involved with observations separated by *s* periods.

Since the model is incorporating both regular and seasonal dependence multiplicatively, the model is called as Multiplicative seasonal ARIMA model. The general model is denoted by SARIMA (p, d, q) $(P,\ D,\ Q)_s$

Where, p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, s = Seasonal Period

SARIMA (p, d, q) $(P,\ D,\ Q)_s$ is written as follow.

$$\emptyset_{\boldsymbol{P}}(\boldsymbol{B^S})\emptyset(\mathbf{B})\nabla_{\boldsymbol{S}}^{\boldsymbol{D}}\nabla^{\boldsymbol{d}}\boldsymbol{X_t} = \boldsymbol{\delta} + \boldsymbol{\theta_Q}(\boldsymbol{B^S})\boldsymbol{\theta}(\boldsymbol{B})\boldsymbol{\omega_t}$$

Where,
ω_t is the usual Gaussian White Noise Process

B = Backward shift operator

$\emptyset(\mathbf{B})$ represents Ordinary Autoregressive component with order p

$\boldsymbol{\theta}(\boldsymbol{B})$ represents Ordinary Moving Average component term with order q

$\emptyset_{\boldsymbol{P}}(\boldsymbol{B^S})$ represents Seasonal Autoregressive component with order P

$\boldsymbol{\theta_Q}(\boldsymbol{B^S})$ represents Seasonal Moving Average component with order Q

$\nabla^d = (1 - B)^d$ represents Ordinary difference component

$\boldsymbol{\nabla_S^D} = (1 - B^s)^D$ represents Seasonal difference component

Moreover, Ordinary Autoregressive component with order p and Ordinary Moving Average component term with order q are written as follows.

Ordinary **AR:** $\emptyset(\mathbf{B}) = \mathbf{1} - \emptyset_{\mathbf{1}}\boldsymbol{B} - \ldots\ldots\ldots - \emptyset_{\boldsymbol{p}}\boldsymbol{B}^{\boldsymbol{p}}$

Ordinary **MA:** $\boldsymbol{\theta}(\boldsymbol{B}) = \mathbf{1} + \boldsymbol{\theta}_{\mathbf{1}}\boldsymbol{B} + \ldots\ldots\ldots + \boldsymbol{\theta}_{\boldsymbol{q}}\boldsymbol{B}^{\boldsymbol{q}}$

Seasonal Autoregressive component with order P and Seasonal Moving Average component with order Q are written as follows.

Seasonal **AR:** $\emptyset_{\boldsymbol{P}}(\boldsymbol{B}^{\boldsymbol{S}}) = \mathbf{1} - \emptyset\, \boldsymbol{B}^{\boldsymbol{s}} - \emptyset\, \boldsymbol{B}^{\mathbf{2}\boldsymbol{s}} - \ldots\ldots\ldots - \emptyset\, \boldsymbol{B}^{\boldsymbol{Ps}}$

Seasonal **MA:** $\boldsymbol{\theta}_{\boldsymbol{Q}}(\boldsymbol{B}^{\boldsymbol{S}}) = \mathbf{1} + \boldsymbol{\theta}_{\mathbf{1}}\boldsymbol{B}^{\boldsymbol{s}} + \boldsymbol{\theta}_{\mathbf{2}}\boldsymbol{B}^{\mathbf{2}\boldsymbol{s}} + \ldots\ldots\ldots + \boldsymbol{\theta}_{\boldsymbol{Q}}\boldsymbol{B}^{\boldsymbol{Qs}}$

2.2.2 Residual Analysis and Model Adequacy

After fitting a model, it is essential to check model adequacy in order to make valid inferences. Generally, checking a model adequacy is based on the residual analysis. Residual Analysis is aimed at finding out whether the underlined distributional assumptions of residuals are validated by the fitted model. Hence, Normal Probability Plot and Anderson Darling Test were carried out for checking normality assumption of residuals.

Moreover, Ljung-Box test was carried out for checking Model Adequacy. Also, ACF of residuals and Augmented Dickey Fuller Test were used for checking independence of residuals and Stationary of the series. Forecasting capacity of the developed model was evaluated using Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE).

3. Results

Initially, time series plot was used for identifying nature of the monthly dengue cases in Sri Lanka. According to the figure 1.3: time series plot, there can be observed jump in the time series after January 2017. Thus some unusual observations can be observed in the series and this was checked using Box plot depicted in the figure 1.4.

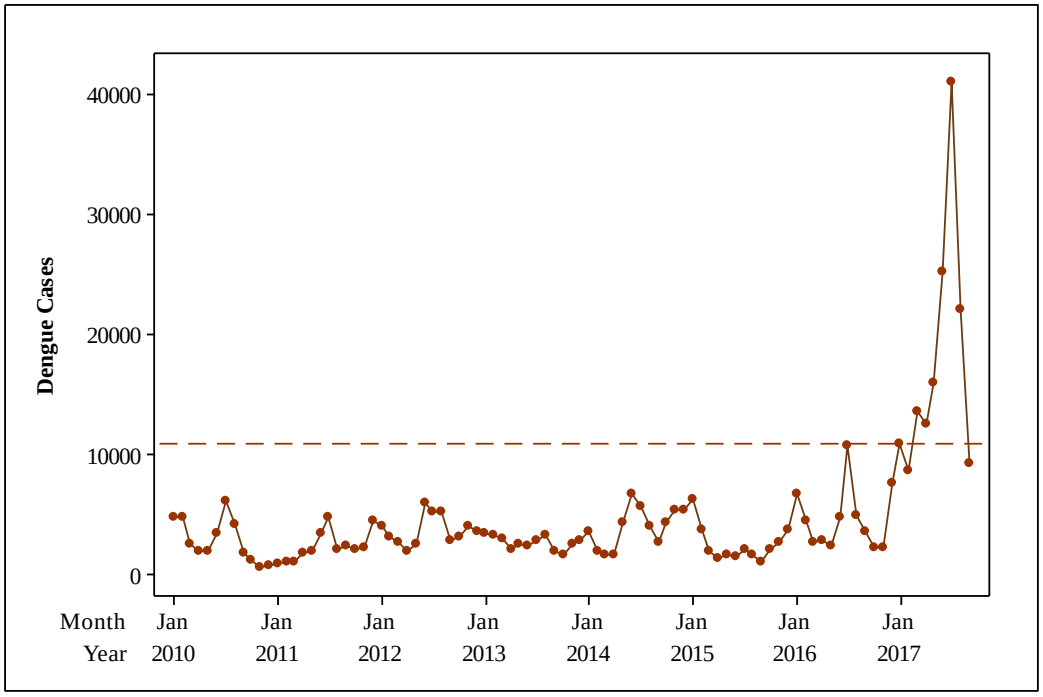

Figure 1.3: Time Series Plot of Monthly Dengue Cases from January 2010 to September 2017 in Sri Lanka

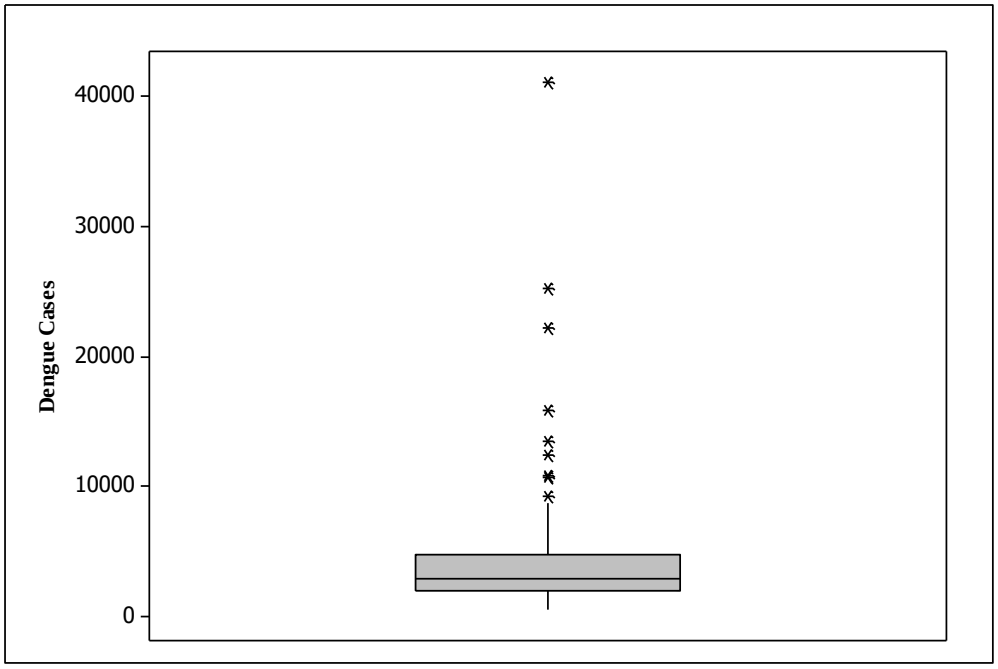

Figure 1.4: Box plot of Monthly Dengue Cases from January 2010 to September 2017 in Sri Lanka

According to Figure 1.4 box plot, outliers can be observed in the series. For the analysis, these outliers have been adjusted using moving average with order three according to the technique used by Konarasinghe, Abeynayake, and Gunaratne (2016).

$$\text{i}^{th} \text{ value of the series } = \frac{[(i-1)^{th}\ value\ + (i-2)\ ^{th} value\ +\ (i-3)^{th}\ value]}{3}$$

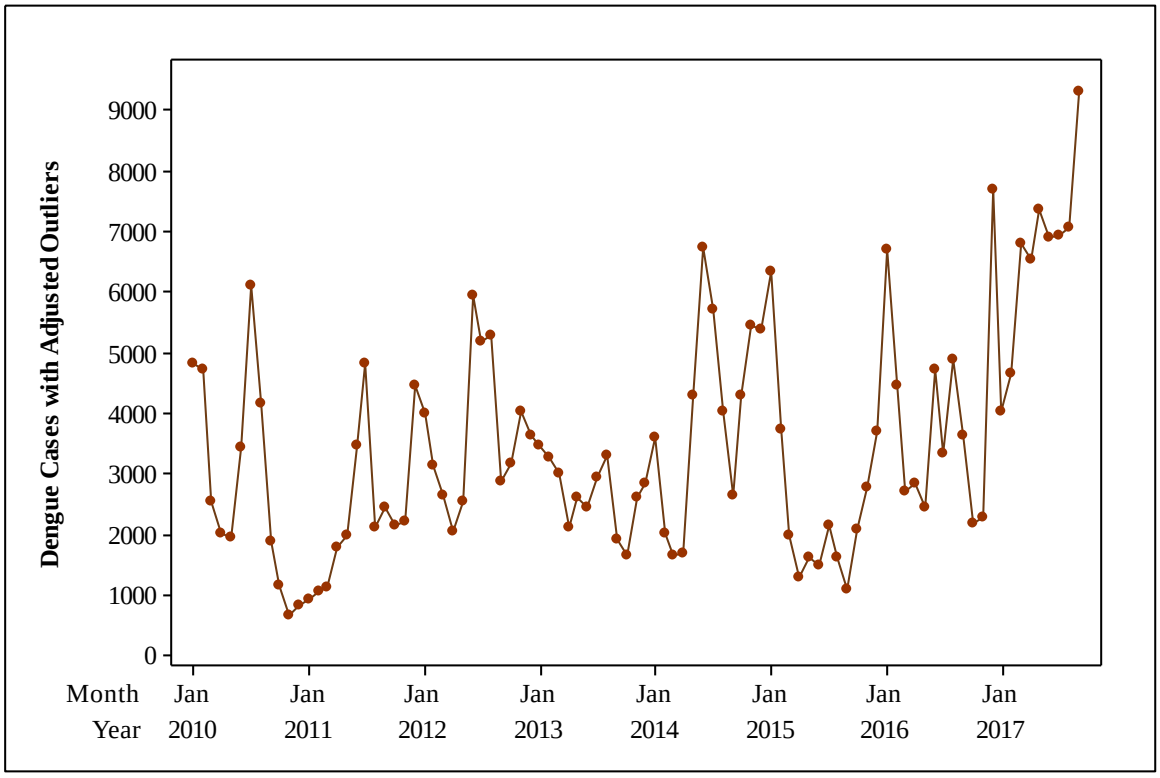

Figure 1.5: Time Series Plot of monthly dengue cases after adjusting outliers

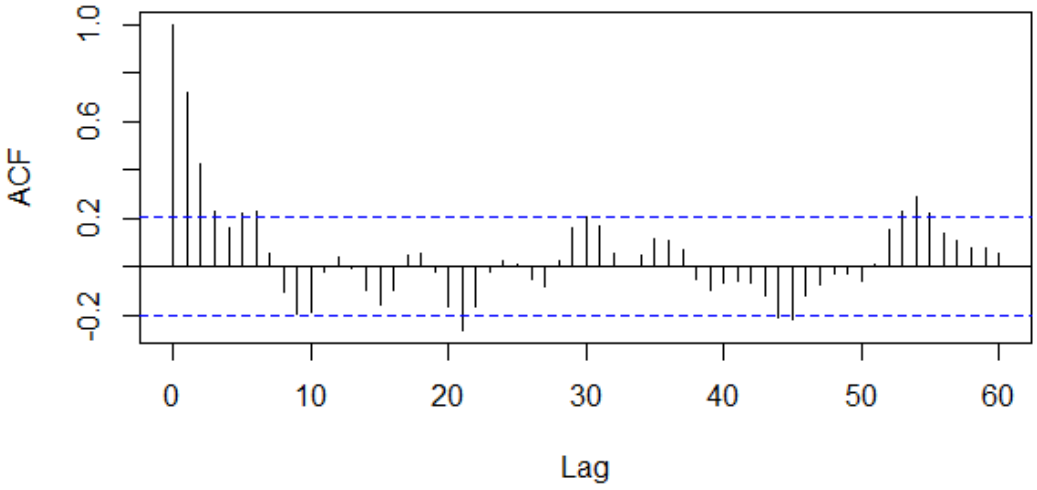

Figure 1.6: ACF of log transformed monthly dengue cases after adjusting outliers

Figure 1.5 depicts the Time Series plot of monthly dengue cases after adjusting outliers and Figure 1.6: ACF plot depicts log transformed monthly dengue cases after adjusting outliers. According to the ACF plot in Figure 1.6, the up and down fluctuations can be observed. Initially, autocorrelations are decreased up to lag 5 and there is an increase of autocorrelation at lag 6. After that autocorrelations have again decreased up to lag 8. Accordingly, there should be a seasonal pattern in the series during 4 months period or 5 months period. Based on the pattern identified by ACF plot in Figure 1.6, SARIMA (p, d, q) $(P,\ D,\ Q)_s$ models have been decided to develop. Here, most appropriate model among developed models was SARIMA(2, 1, 2)(0, 0, 1)$_4$ based on minimum AIC which is 34.54.

Following table illustrates the parameter estimates of the SARIMA(2, 1, 2)(0, 0, 1)$_4$ model

Table 1.1: parameter estimates of the SARIMA(2, 1, 2)(0, 0, 1)$_4$ model

Terms of the model	95% Confidence Interval
Non-seasonal AR(1)	(0.9119615,0.9909131)
Non-seasonal AR(2)	(-1.0152876,-0.9736848)
Non-seasonal MA(1)	(-1.1544711,-0.8958013)
Non-seasonal MA(2)	(0.8719092,1.1280428)
Seasonal MA(1)	(-0.6659237,-0.1242835)

According to above table 1.1, 95% confidence intervals obtained for all coefficients of Non-seasonal AR, MA terms and Seasonal MA term in the fitted model do not include zero. It concludes that the all coefficients of Non-seasonal AR, MA terms and Seasonal MA term in the fitted model are significant.

Residual Analysis for the SARIMA (2, 1, 2) (0, 0, 1)$_4$ model

Table 1.2: Results of Box-Ljung test

Number of Lags	X^2	Degrees of freedom	p-value
12	8.8299	6	0.1834

Following has mentioned the hypothesis of Box-Ljung test and its results.

H_0 : Model is adequate

H_1 : Model is not adequate

According to Table 1.2, since p-value of the test statistics provided by Box-Ljung test is greater than 0.05, H_0 cannot be rejected at 5% level of significance. It concludes that the fitted model is adequate.

Table 1.3: Results of Augmented Dickey-Fuller Test

Dickey-Fuller	Lag order	p-value
-3.9619	3	0.01721

According table 1.3 table, p-value obtained by Augmented Dickey-Fuller Test is less than 0.05 at 5% level of significance. It concludes that the residuals of the fitted model are uncorrelated.

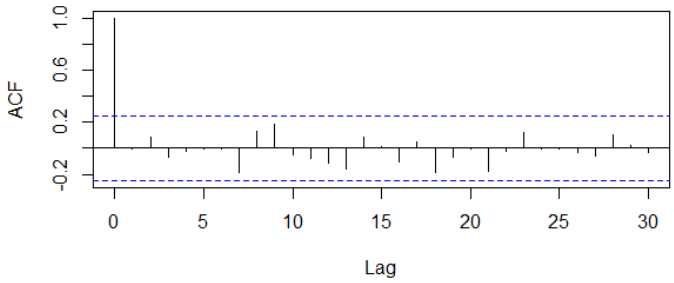

Figure 1.7: ACF plot of residuals of the fitted model

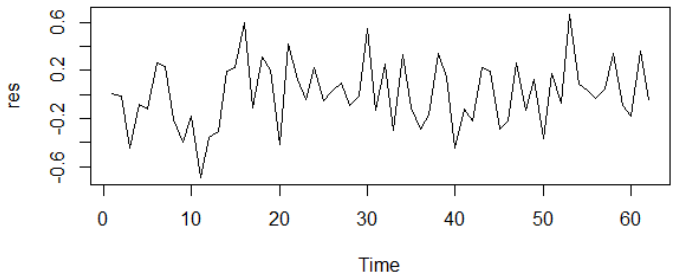

Figure 1.8: Time Series plot of residuals of the fitted model

Moreover, ACF plot and time series plot of residuals are shown by Figure 1.7 and Figure 1.8. According to ACF plot, autocorrelations of all lags except first lag are insignificant.

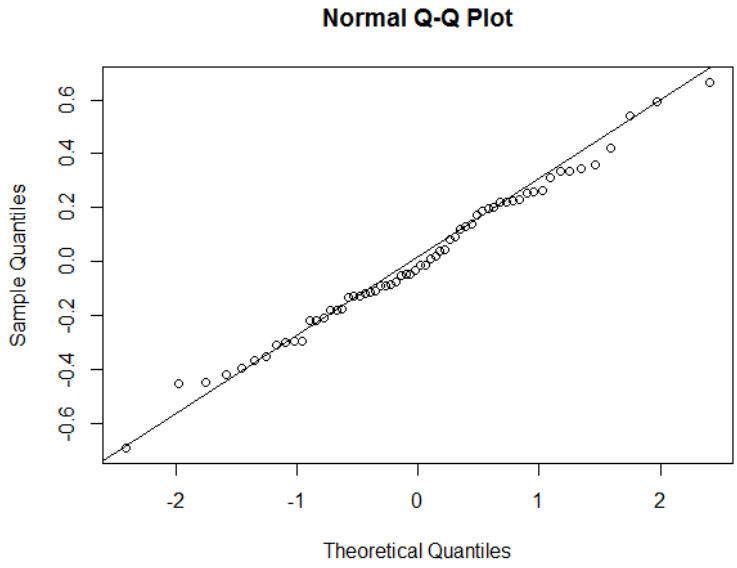

Figure 1.9: Normal Probability plot

Table 1.4: Anderson Darling Test Results

AD Statistics	**0.208**
P-Value	**0.86**

According to Figure 1.9 Normal probability plot, it depicts that the most of the points lie on the line angle at 45^0. Moreover, Anderson Darling Test was carried out. According to Table 1.4, Since P-value of AD statistics (= 0.86) is greater than 0.05, H_o is accepted at 5% level of significance. It concludes that the residuals are normally distributed. According to results of the residual analysis, it is shown that the underlined distributional assumptions of residuals are validated by the fitted model.

Accuracy of the model fitted and verification

Table 1.5: Summary of SARIMA (2, 1, 2) $(0, 0, 1)_4$ model

Type of Model	ME	RMSE	MAE	MPE	MAPE
Model Fitted *(For Fitted and Log transformed Cases with adjusted outliers)*	0.003786	0.273679	0.222321	-0.052040	2.849737
Model Verification *(For Predictions and Log transformed Cases with adjusted outliers)*	-0.244571	0.684696	0.583282	-3.563479	7.405995
Model Verification *(For Predictions and Log transformed Actual Cases)*	0.049809	0.989539	0.854733	-0.649497	10.03829

Abbreviations used: *ME: Mean Error, RMSE: Root Mean Squared Error, MAE: Mean Absolute Error, MPE: Mean Percentage Error, MAPE: Mean Absolute Percentage Error*

According to Table 1.5, it clearly shows that the figures provided by the accuracy measurements for both model fitted and verification with Log transformed Cases with adjusted outliers are significantly small. Moreover, figure 1.10 time series plot depicts that the fitted values and Log transformed Cases with adjusted outliers are very close. According to figure 1.11 time series plot, although predictions are close to Log transformed Cases with adjusted outliers, predictions are underestimates with the Log transformed Actual Cases (without adjusted outliers). It mainly causes due to the substantial increase of the monthly dengue cases after January 2017 (Figure 1.3: Time Series Plot).

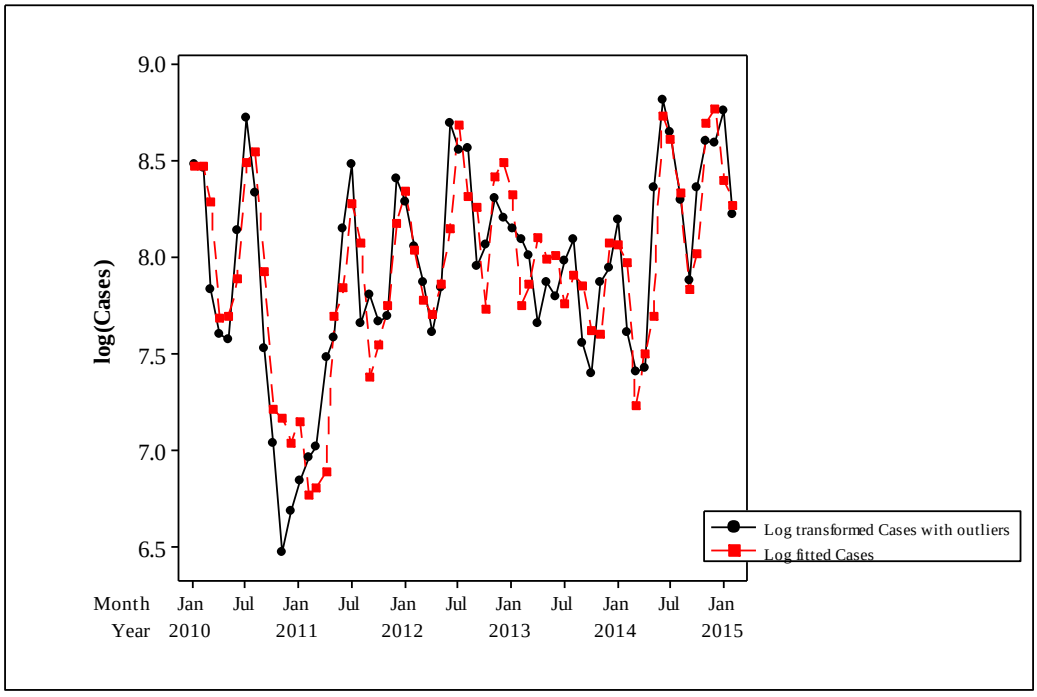

Figure 1.10: Time Series plot of fitted values Vs Log transformed Cases with adjusted outliers

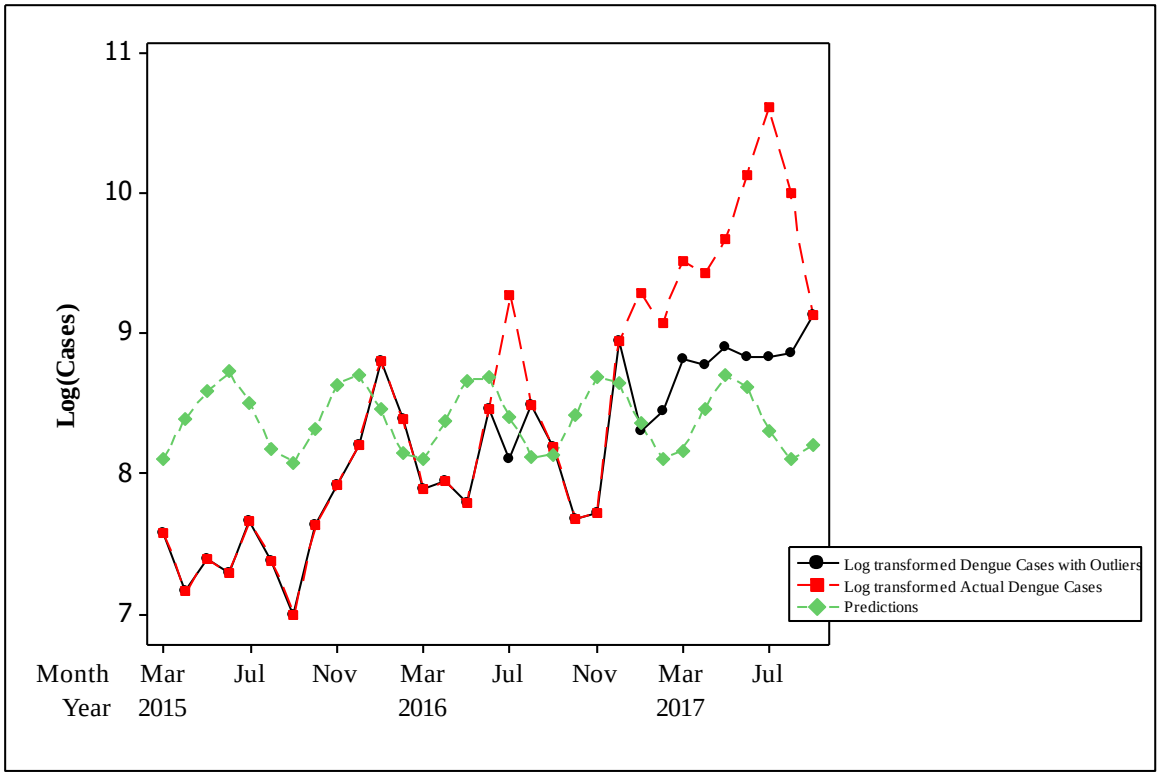

Figure 1.11 Time Series plot of Predictions Vs Log transformed Cases with adjusted outliers and Log transformed Actual Cases

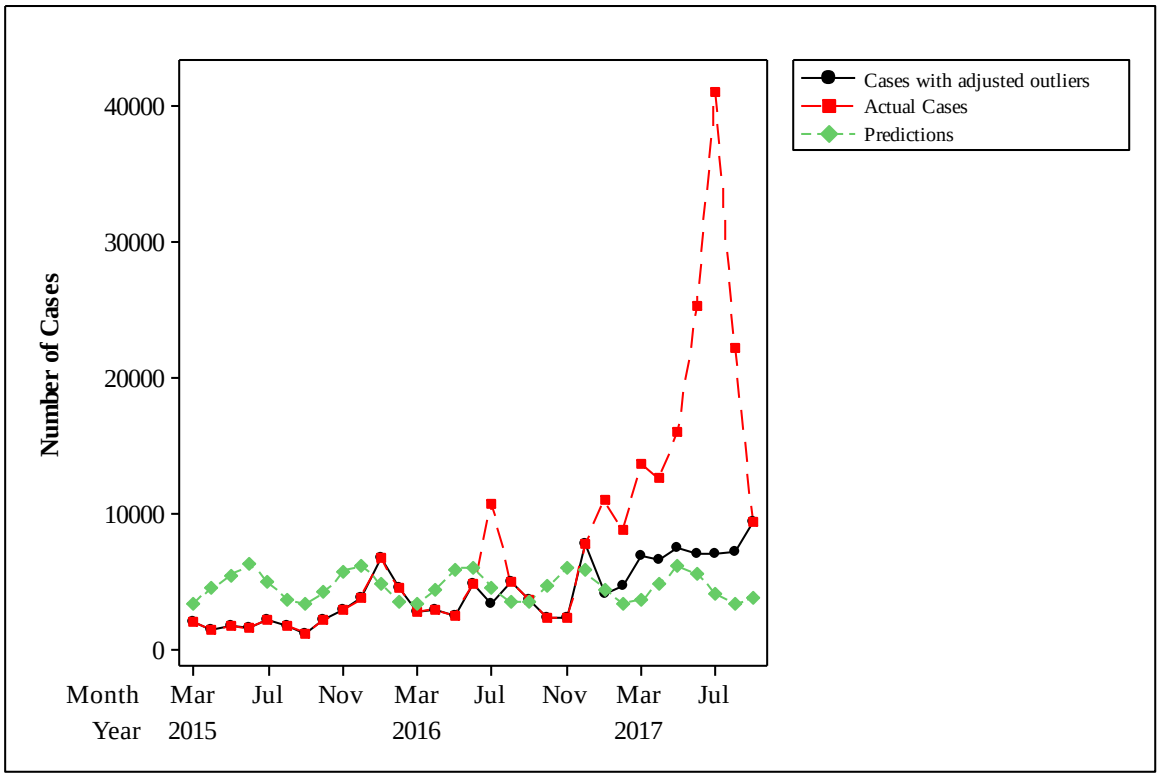

Figure 1.12 Time Series plot of Predictions Vs Cases with adjusted outliers and Actual Cases

Figure 1.12: Time series plot depicts how predictions, number of dengue cases and number of dengue cases with adjusted outliers have changed overtime.

Following has shown all significant coefficients of AR(1), AR(2), MA(1), MA(2) and SMA(1) terms in SARIMA(2, 1, 2)(0, 0, 1)$_4$ model.

Coefficients of the fitted SARIMA (2, 1, 2) (0, 0, 1)$_4$ model

	AR(1)	AR(2)	MA(1)	MA(2)	SMA(1)
Coefficients	0.9514	-0.9945	-1.0251	1.0000	-0.3951
Standard Error	0.0201	0.0106	0.0660	0.0653	0.1382

Table 1.6: Predictions and Prediction Intervals at 95 % level provided by SARIMA (2, 1, 2) (0, 0, 1)$_4$ model

Year	Month	Actual Cases	Cases with adjusted outliers	Predictions	Prediction Interval at 95 % Level	
					Lower	Upper
2015	March	1962	1962	3344	1932	5790
2015	April	1293	1293	4425	2082	9406
2015	May	1625	1625	5377	2208	13094
2015	June	1477	1477	6184	2268	16861
2015	July	2125	2125	4972	1755	14086
2015	August	1604	1604	3579	1198	10689
2015	September	1099	1099	3252	1028	10286
2015	October	2066	2066	4117	1244	13626
2015	November	2762	2762	5667	1660	19341
2015	December	3688	3688	6075	1735	21273
2016	January	6694	6694	4724	1305	17091
2016	February	4439	4439	3469	917	13126
2016	March	2696	2696	3322	838	13173
2016	April	2832	2832	4333	1054	17817
2016	May	2422	2422	5824	1381	24556
2016	June	4731	4731	5926	1374	25553
2016	July	10715	3328	4489	1011	19942
2016	August	4873	4873	3388	733	15660
2016	September	3629	3629	3416	710	16440
2016	October	2185	2185	4556	919	22600
2016	November	2257	2257	5942	1173	30111
2016	December	7677	7677	5746	1111	29721
2017	January	10927	4040	4273	803	22733
2017	February	8724	4658	3334	604	18391
2017	March	13540	6813	3533	618	20192
2017	April	12510	6525	4781	815	28040
2017	May	15936	7353	6016	1006	35955
2017	June	25296	6897	5541	909	33771
2017	July	41026	6925	4078	651	25532
2017	August	22162	7058	3306	511	21403
2017	September	9321	9321	3672	550	24529

Table 1.6 shows prediction values obtained and prediction intervals generated at 95 % level from the developed SARIMA (2, 1, 2) $(0, 0, 1)_4$ model. According to the prediction intervals provided for March 2015 to September 2017, it clearly shows actual number of dengue cases in each month except Month of July in 2017 have included within each corresponding prediction interval.

Discussion and Conclusion:

This study is mainly focused on develop an appropriate time series model for forecast number of dengue cases reported in future months in Sri Lanka. In this study, monthly dengue cases reported from January 2010 to September 2017 has been taken into account for the model development. The developed SARIMA (2, 1, 2) $(0, 0, 1)_4$ model has been selected as most appropriate model among other developed SARIMA (p, d, q) $(P, D, Q)_s$ models for forecasts. The fitted model has satisfied all the model validation criterion and also measurements of errors are satisfactorily small in the fitted model. However, some figures obtained from the model verification have clearly shown that the forecasted values are underestimates, when compared to the actual cases without outlier adjustment. It is due to the jump in the time series depicted in figure 1.3 time series plot. Hence, the fitted model is appropriate to use for forecast dengue cases in Sri Lanka if the number of dengue cases in the future months have not increased remarkably similar to the reported number of dengue cases during the months of January 2017 to August 2017. Moreover, it is observed that the number of dengue cases reported in the month of September 2017 has significantly decreased to the normal level of the time series as depicted in figure 1.3 time series plot. If the future monthly dengue cases do not show unusual observations as previously occurred, the model can be used for predicting dengue cases together with the prediction intervals provided by the model.

Further Research:

Through this study, main focus is to develop a time series predictive model for forecast monthly number of dengue cases in Sri Lanka. Further, it is intended to develop a model for Colombo districts which is observed as the most affected area from dengue disease in Sri Lanka.

Bibliography:

1. Shumway, R.H. and Stoffer, D.S., 2011. Time Series Analysis and its application with R examples.
2. Konarasinghe, U., 2017. Hybrid Trend–ARIMA Model for Forecasting Employment in Tourism Industry in Sri Lanka.
3. World Health Organization 2017, Dengue and severe dengue, viewed April 2017, http://www.who.int/mediacentre/factsheets/fs117/en/