Modeling and Forecasting Dengue Cases in the area of Colombo Municipal Council in Sri Lanka

S. R. Gnanapragasam

Department of Mathematics and Computer Science, The Open University of Sri Lanka

Abstract- Dengue is a mosquito-borne viral disease and it spreads rapidly in many parts of the world including Sri Lanka. The commercial capital of Sri Lanka is Colombo and Colombo Municipal Council (CMC) is the largest and financial centre in Sri Lanka. Reports in health ministry of Sri Lanka say that, more dengue cases are reported in western part of the country, particularly, from areas of CMC. The study mainly focused on the dengue cased reported in the areas of CMC in Sri Lanka. The purpose of forecast is to the plan the future activities. Thus in this study, a statistical model is fitted to predict the future dengue cases in the area of CMC. Then relevant official can take necessary action to control the future dengue cases in CMC. All necessary tests are employed to build the model. Hence the predicted values are obtained. Accordingly, more than 1500 dengue cases are expected in last six months in the year 2016.

Index Terms- Colombo Municipal Council, Dengue Cases, Forecasting, Modeling, Sri Lanka.

I. INTRODUCTION

Dengue is the widespread viral infection in the world today. Now dengue cases are very common in Sri Lanka too. Records in the health ministry of Sri Lanka show that, 215750 total cases are reported in last six years till December 2015. Only during the first 6 months of the year 2016, more than 21700 suspected cases have been reported from all over the island to the epidemiology unit in health ministry of Sri Lanka. [1] showed that over 2.5 billion people infected all over the world in which about 2.5% die.

Since the first reported dengue cases in 1965, there had been reports continuously until the recent past and Sri Lankan population had been exposed to the virus for decades. Now dengue incidences are common in Sri Lanka [6]. In the past, [5] said that, dengue has become the number one killer mosquito infection in Sri Lanka. The meteorological factors which effect on dengue cases in Sri Lanka are discussed in [4] by considering three geographically different district areas of Colombo, Anuradhapura and Ratnapura. The factors in the epidemiological pattern of cases in Sri Lanka are identified in [7]. Recently a statistical model has been fitted in [3] to predict the total number of cases in Sri Lanka. Another study [2] showed that nearly 50% of the cases are from western province of Sri Lanka and further this study reveals that more than 25% of the data are from Colombo district of western province of Sri Lanka.

Colombo district can be further divided into two main parts; one as areas of Colombo Municipal Council (CMC) and the other as out of CMC areas in Colombo district. Colombo district has got 669 km^2 as total land area in which 37 km^2 land of area belongs to CMC. Nevertheless more cases are reported in CMC areas than out of CMC areas. Therefore this study is mainly focused on dengue cases in areas of CMC. The government of Sri Lanka is currently implementing several programmes to control the cases. In addition to these programmes, this study will provide statistical information in terms of forecasting the future cases for the purpose of planning.

Objectives of the study

The aim of this study is to fit a statistical model to forecast the dengue cases in Colombo municipal council areas of Colombo district in Sri Lanka. To achieve this, the following objectives are attained:

- To examine the behavior of dengue cases in island wide, western province, Colombo district and Colombo municipal council in Sri Lanka
- To rank the provinces and districts of Sri Lanka based on the average dengue cases
- To fit time series models to forecast the dengue cases in areas of CMC
- To predict the dengue cases for next 6 months in the year 2016

Based on the forecasted future dengue cases, further planning can be done by the relevant authorities in Sri Lanka.

Source of data

Monthly wise data from January 2010 to June 2016, released by the epidemiology unit of health ministry of Sri Lanka are used for this study. For the purpose of model validation, data from March to June 2016 are used. Meanwhile for model development, data from January 2010 to February 2016 are used.

II. METHODOLOGY

A. At preliminary stage prior to model fittings, the following techniques are carried out.

A1. Time Series Plot: This plot is generally used to get an idea about the data and its behaviour. This is to inspect it for extreme observations, missing data, or elements of non-stationary such as trend or seasonality or cyclic pattern or irregular variations.

A2. Augmented Dickey- Fuller test (ADF): ADF test is used to test whether the series has a unit root. It is to confirm, statistically, that the stationary of series in terms of trend availability. Test statistic for the model $Y_t = \rho Y_{t-1} + u_t$ is

$DF = \frac{\hat{\rho}}{SE(\hat{\rho})} \sim t_{n-1}$, where $-1 < \rho < 1$, u_t is the white noise and n is the number of observations. Hypothesis: H_0: $(|\rho| = 1)$ and series is non-stationary versus H_1: $(|\rho| < 1)$ and series is stationary.

A3. Kruskal- Wallis Test: This test is used to confirm the seasonality in the series. The hypothesis to be tested in this test is H_0: series has no seasonality versus H_1: series has seasonality. The test statistic is defined as: $H = \frac{12}{N(N+1)} \sum_{i=1}^{n_i} \frac{R_i^2}{n_i} - 3(N+1) \sim \chi^2_{L-1}$, where N is the total number of rankings, R_i is the sum of the rankings in a specific season, n_i is the number of rankings in a specific season and L is the length of season.

In time series analysis, a process of examining Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) are to determine the nature of the process under consideration. Graphs of ACF and PACF are obtained to observe the satisfaction of stationary condition.

A4. Autocorrelation function (ACF): ACF at lag k is defined by $\rho_k = \frac{\operatorname{cov}\left[\left(Y_t - \hat{Y}_t\right)\left(Y_{t+k} - \hat{Y}_{t+k}\right)\right]}{\sqrt{\operatorname{var}\left(Y_t - \hat{Y}_t\right)\operatorname{var}\left(Y_{t+k} - \hat{Y}_{t+k}\right)}}$.

The first several autocorrelations are persistently large in the graph of ACF and trailed off to zero rather slowly, it can be assumed that a trend exists and the time series is non-stationary.

A5. Partial autocorrelation function (PACF): PACF between Y_t and Y_{t+k} is the conditional correlation between Y_t and Y_{t+k} and defined as follows: $\phi_{kk} = corr\left(Y_t, Y_{t+k} \mid Y_{t+1}, Y_{t+2}, ..., Y_{t+k-1}\right)$. The PACF between Y_t and Y_{t+k} is the autocorrelation between Y_t and Y_{t+k} after adjusting for Y_{t-1}, Y_{t-2},…, Y_{t-k+1}. Hence for an $\mathrm{AR}(p)$ process the PACF between Y_t and Y_{t+k} for $k > p$ should be equal to zero.

A6. Seasonal differencing method: This method is used to transfer the non-stationary series to stationary series. In this method differences are taken at seasonal lags. If the peaks appear seasonally in the autocorrelation function at particular lags, then it can be assumed that there is a seasonal pattern in the series. It is defined as: $W_t = Y_t - Y_{t-L}$, where L is the length of season.

B. To fit ARIMA model, the following techniques are applied.

B1. Seasonal ARIMA models: A model with combinations of autoregressive terms and moving average terms are generally called as Auto Regressive Moving Averages (ARMA) model. A formulation of an ARMA process is given as: $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - ... - \beta_q \varepsilon_{t-q}$. If the series is not stationary, it can often be converted to a stationary series by differencing. It is generally denoted as $\mathrm{ARIMA}(p, d, q)$, where d indicates the amount of differencing. In some cases, the series shows a repeating or cyclic behaviour. These seasonal patterns can be very effectively used to further improve the forecasting performance. A seasonal ARIMA (SARIMA) model or $\mathrm{ARIMA}(p, d, q)(P, D, Q)_s$ usually contains: Regular $\mathrm{AR}(p)$ and $\mathrm{MA}(q)$ terms that account for the correlation at low lags. Seasonal $\mathrm{AR}(P)$ and Seasonal $\mathrm{MA}(Q)$ terms that account for the correlation at the seasonal lags where d, D and S indicate the amount of regular differencing, seasonal differencing and seasonality respectively.

Model Selection

In time series analysis, sometimes more than one model can fit the data equally well. In this case, numerical criterion such as Akaike information criterion (AIC), Schwartz's Bayesian criterion (SBC) and coefficient of determination (R^2) are used to select the best model. The best model is the one which gives the lowest AIC and SBC values and highest R^2 value.

B2. Akaike Information Criterion (AIC): AIC is often used for model selection. For sample size n, the expression of AIC is given by: $\mathrm{AIC}(k) = n \ln\left(\hat{\sigma}^2\right) + 2k$, where k is the number of parameters in the model and $\hat{\sigma}^2$ is the sample variance of the residuals.

B3. Schwartz's Bayesian Criterion (SBC): SBC is another mostly used technique for model selection in time series analysis. For sample size n, the expression of SBC is given as: $\mathrm{SBC}(k) = n \ln\left(\hat{\sigma}^2\right) + k \ln(n)$, where k is the number of parameters in the model and $\hat{\sigma}^2$ is the sample variance of the residuals.

B4. Coefficient of determination $\left(R^2\right)$**:** R^2 is the proportion of variance of a dependent variable explained by the model. The best model gives the largest R^2 value.

Residual analysis of the fitted model

Before using the model for forecasting, it must be checked for adequacy. Diagnostic checks are performed to determine the adequacy of the model. Thus, the residuals should be random and normally distributed with constant variance.

B5. Normality of residuals: An assumption of ordinary least squares regression analysis is that the errors of a model are normally distributed. The values of Skewness and Kurtosis are considered to check the normality of the residuals. The skewness closer to 0 and the kurtosis closer to 3 suggests the residuals follow a normal distribution.

B6. Durbin-Watson (DW) statistic: The most important test for detecting serial correlation is DW statistic. DW statistic is used to test for randomness of residuals. The test statistic is defined as: $d = \frac{\sum_{t=2}^{n}\left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{t=1}^{n}\hat{u}_t^2}$, where u_t is the white noise of a fitted model. The DW closer to 2 reveals that the residuals have no serial correlation.

B7. Lagrange's Multiplier (LM) test: LM test is used to test the independency of residuals. It is an alternative test of Durbin Watson test for serial correlation among residuals. The null hypothesis to be tested is that, H_0: there is no serial correlation of any order. $W = nR^2 \sim \chi^2_{df}$, where df is the number of regressors in the auxiliary regression (only linear terms of the dependent variable are in the auxiliary regression), R^2 is the determination of coefficients and n is the number of observations.

B8. White's General test: This test is used in order to check constant variance of residuals. Accordingly the null hypothesis is H_0: Homoscedasticity against the alternative hypothesis H_1: Heteroscedasticity. Test statistic is: $W = nR^2 \sim \chi^2_{df}$, where df is the number of regressors in the auxiliary regression (squared terms of the dependent variable are also included in addition to terms in the LM test in auxiliary regression).

B9. Mean Absolute Percentage Error (MAPE): MAPE is used to check the accuracy of the model. It is the average of the sum of the absolute values of the percentage errors. It is generally used for evaluation of the forecast against the validation sample. To compare the average forecast accuracy of different models, MAPE statistic is used and it is defined as follows: $\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100$.

III. PRELIMINARY RESULTS AND DISCUSSIONS

In this preliminary analysis part, the behavior of the data is described. It includes the descriptive statistics of the data of island wide records, western province records, Colombo district records and Colombo municipal council records. For this purpose the data from January 2010 to December 2015 are taken for the analysis.

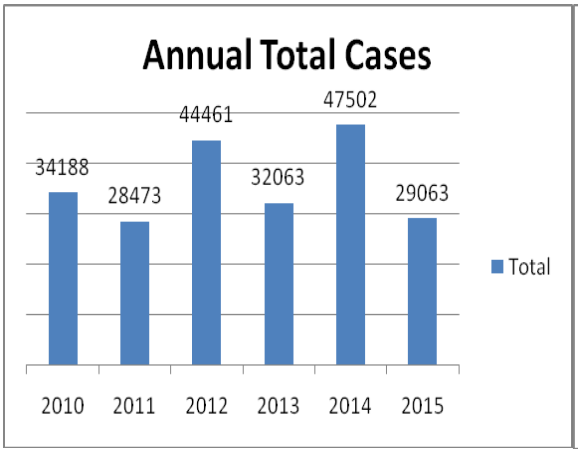

Figure 1(a): Chart of annual total cases

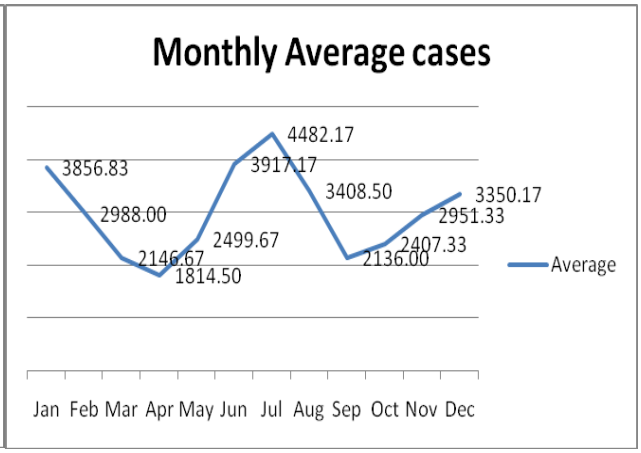

Figure 1(b): Plot of average monthly cases

It can be obtained from Figure 1(a) that, the total number of cases in last six years is 215750 and hence 99.88cases per day are reported island wide. The most number of cases is recorded in the year 2014 while least number of cases is recorded in the year 2011. It can be clearly seen a pattern that the cases in odd years are less than the previous even year.

From Figure 1(b), it can be clearly observed on the average that, in the months of March, April and September the least number of cases are recorded while in the months of January, June, July and December the cases are high on the average. Therefore it can be claimed that in these two seasons, December-January and June- July, the dengue spread all over the country in

high rates. Table 1 gives the rank of provinces of Sri Lanka on dengue cases.

Table 1: Province rank on dengue cases in Sri Lanka

Rank	Province	Percentage	Rank	Province	Percentage	Rank	Province	Percentage
1	Western	49.26%	4	Central	7.38%	7	Northern	5.75%
2	Sabaragamuwa	10.16%	5	Southern	6.89%	8	Uwa	3.29%
3	North Western	8.55%	6	Eastern	6.05%	9	North Central	2.63%

It is noted from Table 1 that nearly 50% of total cases in Sri Lanka are reported from the western province. The other ranks of provinces are as appeared in Table 1. Table 2 provides the rank of districts in Sri Lanka on dengue cases. The percentage gap between 1st and 2nd ranks of provinces is nearly 40%.

Table 2: District rank on dengue cases in Sri Lanka

Rank	District	Percentage	Rank	District	Percentage	Rank	District	Percentage
1	Colombo	28.41%	10	Batticaloa	3.08%	19	Moneragala	1.15%
2	Gampaha	15.25%	11	Puttalam	2.77%	20	Polonnaruwa	1.07%
3	Ratnapura	6.26%	12	Matara	2.29%	21	N Eliya	0.73%
4	Kurunegala	5.78%	13	Badulla	2.14%	22	Mannar	0.64%
5	Kalutara	5.60%	14	Ampara	1.68%	23	Vavuniya	0.55%
6	Kandy	5.19%	15	Anuradhapura	1.62%	24	Mulativu	0.22%
7	Jaffna	4.13%	16	Hambantota	1.50%	25	Kilinochchi	0.21%
8	Kegalle	3.87%	17	Matale	1.46%			
9	Galle	3.10%	18	Trincomalee	1.29%			

According to the rank of district on cases in Table 2, the highest number of cases are reported in Colombo district and it is noted that over one forth of total cases are recorded from Colombo district. The rest of the ranks of districts are as appeared in Table 2. The difference in percentage between 1st and 2nd districts rank is more than 13%.

Dengue cases in western province of Sri Lanka

The pie chart below gives the dengue cases only in the districts, Colombo, Gampaha and Kalutara, of western province of Sri Lanka.

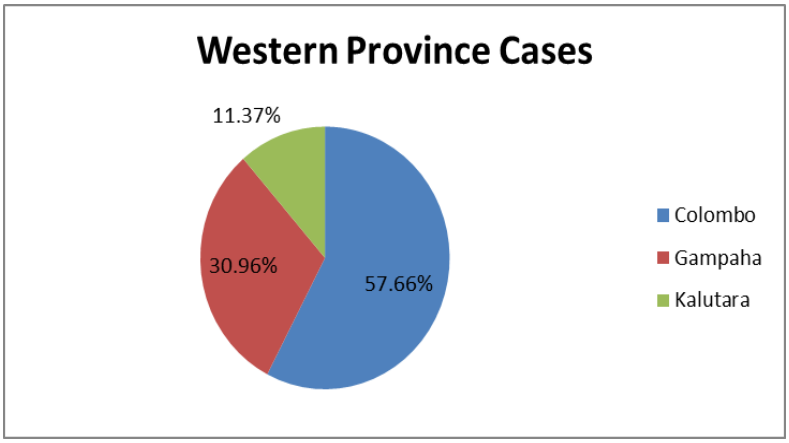

Figure 2: Pie chart of cases in western province

It can be clearly seen from Figure 2 that, more dengue cases (57.66%) are recorded from Colombo District while 30.96% from Gampaha district and 11.37% cases are from Kalutara district in western province of Sri Lanka. Hence it can be stated that, more than the half of the cases in western province are recorded in Colombo district.

Dengue cases in Colombo district of Sri Lanka

The land area belongs to CMC is only 37 km^2 and the rest is nearly 662 km^2 area in Colombo district. However, more cases are recorded in areas of CMC than other areas in Colombo district. When the split data are concerned, as cases in areas of CMC and cases in other areas of Colombo district, 29.25% of cases are recorded in areas of CMC which is 8.31% of total number of cases in Sri Lanka. Thus it is actually third rank, if it is compared with Table 2 of district rank on cases. The statistics relevant to areas in CMC are summarized in Table 3 below:

Table 3: Statistics on dengue cases in areas of CMC

Year	Total Cases	Monthly Average	Cases per day
2010	1924	160.33	5.27
2011	3224	268.67	8.83
2012	2611	217.58	7.13
2013	3861	321.75	10.58
2014	3618	301.50	9.91
2015	2693	224.42	7.38
Total	**17931**	**Grand mean**	**8.30**

As per the statistics appeared in Table 3, 17931 total cases are reported from the year 2010 to 2015. It is noted that, the lowest cases are reported in 2010 with 2.27 cases per day. Particularly in the year 2013, most cases are reported with 10.58 cases per day. It is the highest in the recent past in areas of CMC. At the same time, on the average in last six years 8.30 cases per day are reported in areas of CMC. It can be concluded that most of the cases are recorded from the western part of the country and in which more cases are reported from Colombo district particularly from the areas of CMC.

IV. DEVELOPING ARIMA MODEL FOR DENGUE CASES IN CMC AREAS

In this section, the data from January 2010 to February 2016 are used to build ARIMA model. Thus the monthly records during this period in the areas of CMC are taken for this analysis.

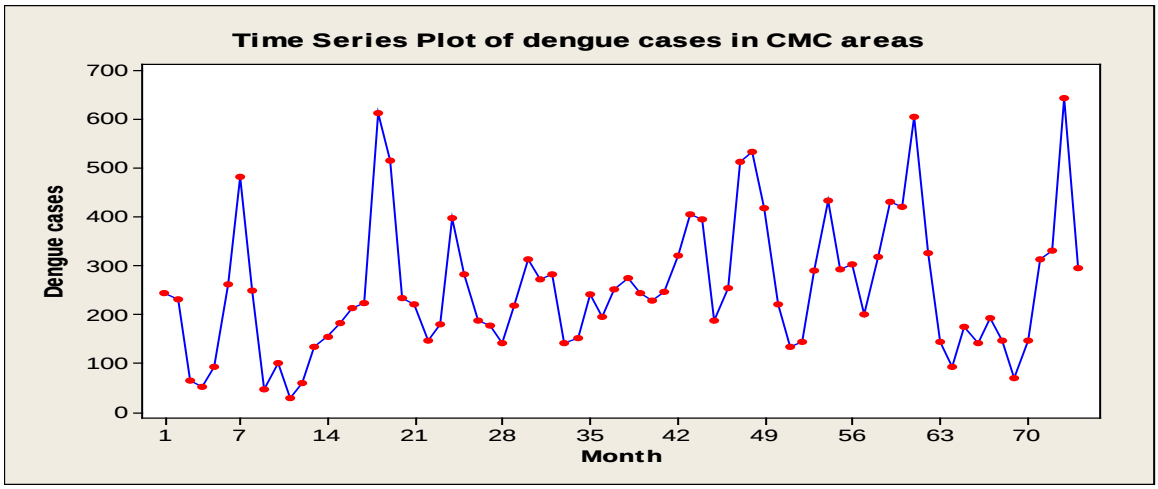

Figure 3: Time series plot of dengue cases in areas of CMC

It seems from the time series plot in Figure 3 that, the series may have the trend and also there can be a seasonal pattern as well. To verify these, ADF test and Kruskal- Wallis test are carried out. According to p-value (0.00) of ADF test, it can be concluded with 95% confidence that the series has no trend. However, the p-value (0.012) of the Kruskal- Wallis test confirms that the series has seasonality. Therefore, it can be concluded that the original series is non-stationary. (It is noted that, for the convenient of software uses, the original data for the dengue cases in CMC areas is named as Y).

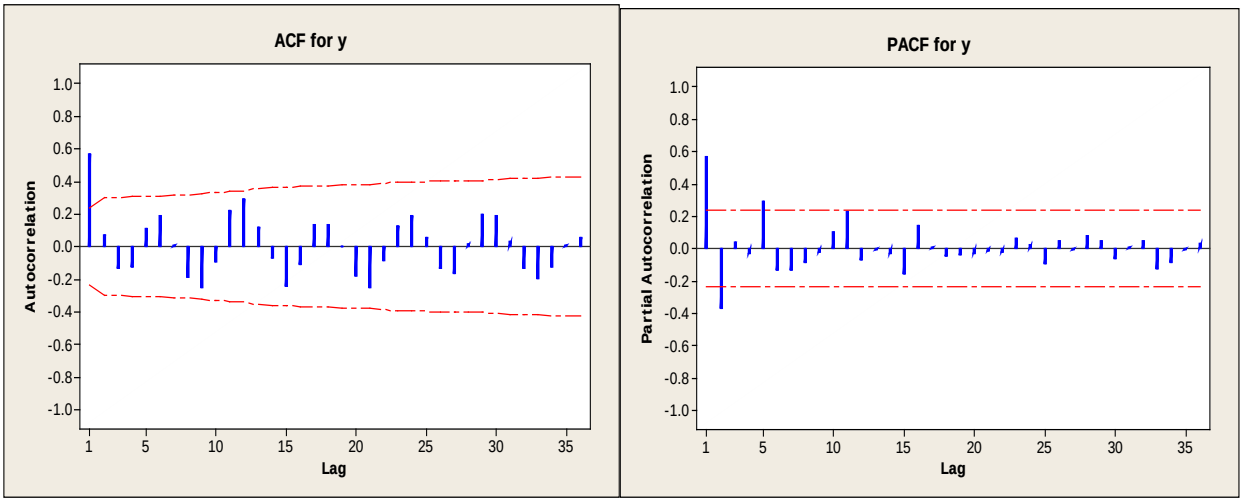

Figure 4: ACF and PACF graphs of the original series

ACF graph in Figure 4 is not decaying exponentially with lags and there are significant spikes in PACF graph. These also confirm that the series is non-stationary. Further, it can be very clearly seen from ACF graph that, there are high spikes at every 6^{th} lags. This pattern suggests that the series can have the seasonality 6. Thus the 6^{th} difference is taken to the original series to remove this seasonal pattern. (Here the new data obtained after 6^{th} differenced is named as D6Y in the analyzing part)

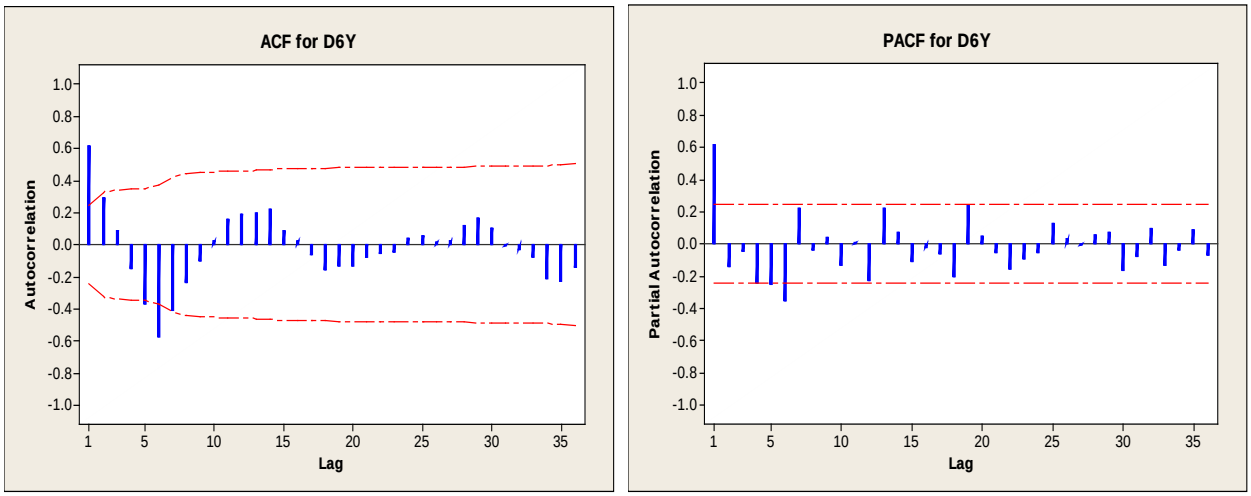

Figure 5: ACF and PACF graphs of the series with 6^{th} difference

Once again the ADF and Kruskal- Wallis tests are carried out to check the stationarity. According to p-value (0.00) of ADF test, it can be concluded with 95% confidence that the series has no trend. Also, the p-value (0.189) of the Kruskal- Wallis test confirms that the series has no seasonality. Therefore, it can be concluded that the series with 6^{th} difference is stationary. Hence this series can be used for seasonal ARIMA model development.

The appearance of ACF and PACF graphs in Figure 5 are also confirms that the series with 6^{th} difference is stationary. Moreover, the model may contains MA(1), MA(5) and SMA(1) as 1^{st}, 5^{th} and 6^{th} terms are significant in ACF graph. At the same time, AR(1), AR(4), AR(5), SAR(1) and SAR(2) terms may include in the model as 1^{st}, 4^{th}, 5^{th}, 6^{th} and 12^{th} spikes are significant in PACF graph in Figure 5. Hence, all possible combinations of these terms are taken in to account to build the models. In which, the most appropriate three models are summarized in Table 4.

Table 4: Test results of diagnostic checking of selected models

Model	DW	Skewness	Kurtosis	LM	White General
ARIMA (1, 0, 1) (1, 1, 0)$_6$	2.04	0.20	3.08	0.70	0.90
ARIMA (1, 0, 0) (1, 1, 2)$_6$	1.93	0.28	3.13	1.00	0.12
ARIMA (1, 0, 5) (0, 1, 0)$_6$	1.89	0.17	2.87	0.66	0.20

As per the test statistics appeared in Table 4, the DW values of all three model are closer to 2 indicates that residuals in all the models are not serially correlated. Also p- values of White general test of all the models in Table 4 are not significant. That reveals the residuals of all three models have constant variance.

Skewness and kurtosis values of all three models are closer to 0 and 3 respectively. Hence it can be confirmed that, the residuals in these models are normally distributed. Also the p-values of LM test are not significant. Thus it can be concluded with 95% confidence that, the residuals of all the models in Table 4 are independent. Hence it can be stated that these three models satisfy all necessary conditions in diagnostic checking.

Since all three models are adequate, the selection criterions are taken to select the best model for forecasting. Thus the test results of selection criteria of all three models are reported in Table 5.

Table 5: Test results of selection criteria

Model No	Model	R^2	AIC	SBC	MAPE	Rank
1	ARIMA (1, 0, 1) $(1, 1, 0)_6$	66 %	12.18	12.32	20.44	3
2	ARIMA (1, 0, 0) $(1, 1, 2)_6$	72 %	11.95	12.06	**14.72**	1
3	ARIMA (1, 0, 5) $(0, 1, 0)_6$	67 %	12.11	12.21	26.63	2

Based on the test results of selection criteria in Table 5, model 2 has largest R^2 value and lowest AIC and SBC values. Therefore model 2 is tentatively selected as best model. On the other hand, ex-ante forecast is calculated for all three models. Here as well model 2 has the least MAPE value. Thus model 2, **ARIMA (1, 0, 0) $(1, 1, 2)_6$**, can be selected as the best model.

Hence the equation of the best model is:

$$Y_t = 0.34 * Y_{t-1} + 0.27 * Y_{t-6} - 0.84 * e_{t-12} + e_t$$

where Y_t is the dengue cases at time t, Y_{t-1} and Y_{t-6} are preceding cases at time $t-1$ and $t-6$ respectively, e_{t-12} is residual at twelve preceding period $t-12$, e_t is the residual at time t.

Since the selected model satisfies all important condition of diagnostic testing, it can be concluded that, the selected seasonal ARIMA model is significant and it can be used for forecast the dengue cases in future. Table 6 below gives the monthly wise predicted dengue cases (with its 95% confidence level) from the month of July to December in 2016.

Table 6: Ex-post forecast of dengue cases

Month in 2016	Ex-post Forecast	95% Confidence level Lower	Upper
July	329	233	524
August	262	141	483
September	157	117	386
October	187	143	417
November	302	171	532
December	345	214	576
Total cases	**1,582**	**1,019**	**2,918**

As per the predicted dengue cases appeared in Table 6, it can be expected that 1582 dengue cases will be reported in the areas of CMC for the last 6 months period in the year 2016. Further it can be concluded with 95% confidence that the estimated dengue cases will be in between 1,019 and 2,918.

V. CONCLUSION

It is concluded in the period from January 2010 to December 2015, the total dengue cases reported in Sri Lanka is 215750 and hence nearly 100 cases per day are recorded island wide. On the average, December- January and June- July are the two seasons where the dengue fever spreads all over the country in high rates.

This study further concludes nearly 50% of the total dengue cases are reported in western province of Sri Lanka and one fourth of total dengue cases are from Colombo district. Thus western part of the country is mostly effected area in Sri Lanka. Moreover the study reveals that, nearly one third of dengue cases in Colombo district are from CMC areas. Besides more than 8 dengue cases per day are reported in CMC areas.

It can be concluded that, the fitted time series model to forecast dengue cases in CMC areas is $\hat{Y}_t = 0.34 * Y_{t-1} + 0.27 * Y_{t-6} - 0.84 * e_{t-12}$ with MAPE value

less than 15%. Over 1500 dengue cases are expected from next 6 months in the year 2016 and it is 31.5% increase from the dengue cases reported in last 6 months in the year 2015.

REFERENCES

[1] E. Arul, B. T. Say, W. S. Annelies and M. David, "Comparing Statistical Models to Predict Dengue Fever Notifications," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 758674, 6 pages. doi:10.1155/2012/758674, 2012.

[2] S. R. Gnanapragasam, T. M. J. A. Cooray , "Time Series Models to Forecast Dengue Fever Incidences in Western Province of Sri Lanka," *Proceedings of 8th International Research Conference, General Sir John Kotelawala Defense University Sri Lanka, 2015,* pp. 46-52

[3] S. R. Gnanapragasam, T. M. J. A. Cooray, "Prediction of dengue fever cases in Sri Lanka using time series model," *Proceedings of 3rd Ruhuna International Science and Technology Conference,* 2016, pp. 35

[4] K. Goto, B. Kumarendran, S. Mettananda, D. Gunasekara, Y. Fujii and S. Kaneko, "Analysis of Effects of Meteorological Factors on Dengue Incidence in Sri Lanka Using Time Series Data," *PLoS ONE* 8(5): e63717. doi:10.1371/journal.pone.0063717,2013.

[5] G. N. Malavige, N. Fernando and G. Ogg, "Pathogenesis of Dengue viral infections: *Sri Lankan Journal of Infectious Diseases*," Vol 1(1), 2011, pp. 2-8.

[6] National Plan of Action Prevention and Control of Dengue Fever 2005-2009, *Epidemiology Unit, Ministry of Health Sri Lanka* , 2010, Available http://www.epid.gov.lk/web/images/pdf/Circulars/latest_draft_poa_for_dfdhf.pdf [Accessed: 30th June 2016]

[7] P. D. N. N. Sirisena and F. Noordeen, "Evolution of dengue in Sri Lanka-changes in the virus, vector, and climate," *International Journal of Infectious Diseases*, Vol 19, 2014, pp. 6-12.

AUTHORS

First Author – S. R. Gnanapragasam, Department of Mathematics and Computer Science, The Open University of Sri Lanka, srgna@ou.ac.lk