Time Series Models to Forecast Dengue Fever Incidences in Western Province of Sri Lanka

SR Gnanapragasam[1#] and TMJA Cooray[2]
[1]*Department of Mathematics and Computer Science, The Open University of Sri Lanka, Nugegoda, Sri Lanka*
[2]*Department of Mathematics, University of Moratuwa, Moratuwa, Sri Lanka*
[#]srgna@ou.ac.lk

Abstract*— Dengue is one of the most dangerous mosquito viral infections in the world. In the recent past dengue has become the number one killer mosquito infection in Sri Lanka. Each year the number of incidences of dengue is dramatically increasing. According to the records in Health Ministry of Sri Lanka, 47246 dengue fever incidences are reported in 2014. Since awareness of dengue infections is of utmost importance, Centre for Dengue Research at the Department of Microbiology of Faculty of Medical Sciences of University of Sri Jayewardenepura has been established in 2012 by the government of Sri Lanka. The descriptive statistics of district wise data indicate that more cases are reported from Colombo district in Western province. Therefore the objective of this study is to fit models for Colombo district in particular and Western province of Sri Lanka in general. Accordingly, national level arrangements can be implemented by relevant body to control the incidences as well as to take necessary action in preparation for treatments. The relevant tests in time series analysis are carried out to develop two ARMA models, one for Western Province, and one for Colombo district. The data, from January 2010 to December 2014, published by Health Ministry of Sri Lanka are taken to develop models. The fitted models are used to forecast next-three months incidences from January to March 2015. To check the validation of models Augmented Dickey-Fuller test, Box-pierce Q statistic, serial correlation test and White heterosedasticity test are used. MAPE statistics is used to get accuracy of fitted models. Through this study, it can be confirmed that 50% of total incidences are reported in Western province of Sri Lanka. Further, it is confirmed that one fourth of the total incidences are from Colombo district in Western province. MAPE statistics show nearly 15% and thus the fitted models are acceptable for forecast the incidences.*

Keywords*— Dengue, Forecast, Western*

I. INTRODUCTION

Dengue viral infection is one of the most dangerous mosquito viral infections around the world. The number of incidences has dramatically increased in recent part years and Arul*et al*. (2012) have shown that over 2.5 billion people are infected now in which about 2.5% die. Although the Sri Lankan population had been exposed to the virus for decades, severe forms of dengue infection were rare. But based on the report national plan (2010), nowadays dengue incidences are common in Sri Lanka. Since the first reported outbreak of dengue fever in 1965, there had been outbreaks on and off until the recent past with progressively large outbreaks occurring more frequently.

Malavige *et al*. (2011) have shown that, in recent past years dengue has become the number one killer mosquito infection in Sri Lanka. Weekly Epidemiological Report (2009)showed that, the number of incidences of dengue appears to be rising each year due to several factors. Now in Sri Lanka dengue incidences are reported weekly from all over the Island by Health Ministry of Sri Lanka. Accordingly, 47246 incidences are reported in 2014 year. The distribution of notification dengue cases by month in epidemiology unit (2015) shows that, the dengue peak was observed during the months of May and June in 2012. Further it indicates that the Western province of the country is most affected part of the Island in this regard.

Gota *et al*. (2013) have deeply discusses the meteorological factors which effects on dengue incidences in Sri Lanka by considering three geographically different areas of the island which included the mostly effected Colombo district. Sirisena and Noordeen (2014) have identified the factors in the epidemiological pattern of incidences. Through this it is further recommended to implement effective vector control programmes in the country to reduce the morbidity and mortality associated with the incidence. Therefore this study will help to implement such programmes in terms of forecasting the incidences. Murugananthan *et al*. (2014) have argued that incidences showed a seasonal variation in the distribution incidences in Northern Province. It is identified that the absence of laboratory diagnosis as the major drawback noted in their study. However in terms of laboratory diagnosis the

Western province is much better than that of in Northern province.

It is evident that, special second mosquito control programme (2015), through national dengue control unit and epidemiology unit along with the stake holders of Presidential task on Dengue Prevention, the government of Sri Lanka is implementing several programmes to control the incidences, in addition to the local governing authority activities. Because of the importance of awareness of the dengue infections, Centre for Dengue Research (2012) at the Department of Microbiology of the Faculty of Medical Sciences of the University of Sri Jayewardenepura has established in 2012. The researchers are encouraged to carry out research on dengue by the government through CRD. More findings and the reports of Epidemiology Unit, Ministry of Health, Sri Lanka show that more incidences reported in Western province specifically in Colombo district. This study is also confirmed statistically that approximately 50% of the incidences are reported from Western province while one fourth of total incidences are from Colombo district.

A. *Objective of the study*

The objective of this study is to develop time series models to forecast the incidences in most affected areas in Sri Lanka such as Western province particularly in Colombo district. Accordingly the arrangements to treat them and save their lives can be done by the relevant authority. Further it can be given directions to the relevant authority for more attention to control for the areas which has more dengue patients.

B. *Source of data*

Five years data, monthly wise, released by the Epidemiology Unit of Health Ministry of the Sri Lanka from 2010 January to 2014 December is used for this study. For the purpose of forecasting the data from January 2015 to March 2015 is used. EViews and MINITAB software are used for analysis.

II. METHODOLOGY

Time series analysis is used to fit the Auto Regressive Moving Averages (ARMA) models and forecast for three points ahead.

A. Hypotheses test for proportion

The Hypothesis $H_0 : \rho \leq \rho_0 \;\; VS \;\; H_1 : \rho > \rho_0$ is used to claim that the particular percentages of the incidences are recorded in particular places while the descriptive statistics are obtained to observe them.

B. *Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF)*

Graphs of ACF and PACF are obtained to see the stationary as well as to guess the terms involved in ARMA models.

1) Autocorrelation function (ACF):

Autocorrelation function at lag k is defined by

$$\rho_k = \frac{\operatorname{cov}\left[\left(Y_t - \hat{Y}_t\right)\left(Y_{t+k} - \hat{Y}_{t+k}\right)\right]}{\sqrt{\operatorname{var}\left(Y_t - \hat{Y}_t\right)\operatorname{var}\left(Y_{t+k} - \hat{Y}_{t+k}\right)}}$$

2) Partial autocorrelation function (PACF):

Partial autocorrelation function between Y_t and Y_{t+k} is the conditional correlation between Y_t and Y_{t+k} and defined as follows

$$\phi_{kk} = Corr\left(Y_t \, , \, Y_{t+k} \mid Y_{t+1}, \, Y_{t+2}, \ldots, Y_{t+k-1}\right)$$

C. *The General Mixed* $\mathrm{ARMA}(p,q)$ *Process*

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \beta_1 e_{t-1} + \beta_2 e_{t-2}$$
$$+ \ldots + \beta_q e_{t-q} + e_t$$

Where e_t is white noise and e_t □ $iidN\left(0, \, \sigma^2\right)$.

When building the Time series models some assumptions on the error terms are made. The assumptions are: the error terms should be normal and independent with constant variance. It is necessary to do the diagnostic tests to check whether the assumptions are satisfied by the fitted models.

D. *Normality of error terms*

Skewness and Kurtosis are used to check the normality of the error terms. The skewness closer to 3 and the kurtosis closer to 0 suggests the error terms follow normal distribution.

E. *Lagrange's Multiplier (LM) test*

Lagrange's Multiplier (LM) test is used to test the serial correlation among error terms. The null hypothesis to be tested is that, there is no serial correlation of any order.

F. *White's General test*

White's General test is used in order to check constant variance of error terms. Accordingly the null hypothesis is H0: Homoscedasticity against the alternative hypothesis H1: Heteroscedasticity.

To check the validation of models Augmented Dickey-Fuller test Box-pierce Q statistic, serial correlation test and White heterosedasticity test are used.

G. Augmented Dickey- Fuller (ADF) test

ADF is used to test whether the series has a unit root.

H. Ljung and Box Q statistic

Ljung and Box Q statistic is used to test the hypothesis that the residuals are independently distributed using the statistics: $Q = n(n+2)\sum_{j=1}^{k}(n-j)^{-1}\hat{\rho}_j^2$

I. Coefficient of determination R^2

The coefficient of determination R^2 is used to evaluate the goodness of fitted model. The best model gives the largest R^2 value.

J. Durbin-Watson (DW) statistic

Durbin-Watson (DW) statistic is used to test for randomness of error terms. The DW closer to 2 reveals that the error terms are randomly scatted.

K. Akaike Information Criteria (AIC) and the Schwartz's Bayesian Criterion (SBC)

The Akaike Information Criteria (AIC) and the Schwartz's Bayesian Criterion (SBC) are used to select the best model.

1) *Akaike Information Criteria:* $AIC(k) = n\ln\left(\hat{\sigma}^2\right) + 2k$
2) *Schwartz's Bayesian Criterion:* $SBC(k) = n\ln\left(\hat{\sigma}^2\right) + k\ln(n)$

The best model is the one which gives the lowest AIC and SBC values.

L. *Mean Absolute Percentage Error (MAPE)*

Mean Absolute Percentage Error (MAPE) statistics used to check the accuracy of the fitted models.

$$MAPE = \sum_{t=1}^{n}\left|\frac{e_t}{n}\right| \times 100$$

If MAPE is less than 10% then the fitted model is excellent but if it is less than 15% is a better model. However MAPE less than 20% is acceptable.

III. RESULTS AND DISCUSSION

The results and its interpretations are presented as two different cases one for Western province and the other one for Colombo district.

A. Results of Western province

According to the descriptive statistics, average incidences recorded in Western province is 1497 which is 48.73% of total incidences. Hence it can be claimed that 50% of the total incidences are reported in Western province. To confirm this result the following hypothesis test is performed.

$$H_0 : \rho \le 0.5 \quad VS \quad H_1 : \rho > 0.5$$

Test Statistic Z= -0.20>-1.64 (table Z value at 5% level of significance)

Based on the above hypothesis testing, it can be confirmed that proportion of incidences in Western province is greater than 0.5. Thus it can be concluded with 95% confidence that 50% of total incidences are reported in Western province.

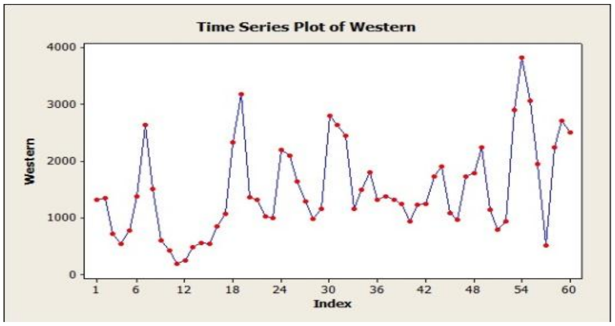

Figure1. Time series plot of western province

As per the Figure 1, it can be seen that, there is no seasonal or cyclic pattern. But it seems that there may be a trend. Thus ADF test is used to check whether there is a trend in the series.

Table 1.ADF Test statistics for Western province series

	t-Statistics
ADF test statistics	-5.025535
Test critical values: 1% level	-3.548208
5% level	-2.912631
10% level	-2.594027

By comparing t- values of the ADF test statistic in all three significance levels as appear in Table 1, it can be confirmed with 99% confidence that there is no trend in the Western province series.

The following two figures Figure 2 and Figure 3 are obtained through MINITAB to check the stationary condition as well as to guess the terms involved in the ARMA models.

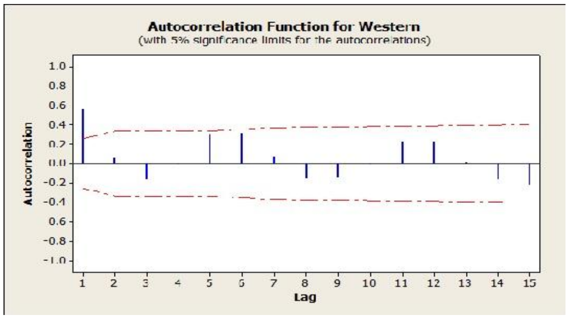

Figure 2. ACF graph for Western province series

From ACF graph in Figure 2, it can be observed that the model may containMA(1) and MA(5) terms as the corresponding spikes at lag 1 and lag 5 are significant in ACF graph for the Western province series.

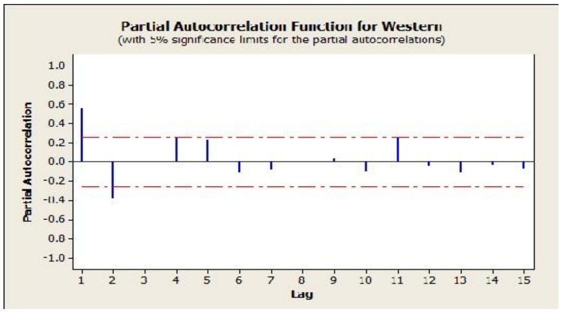

Figure 3. PACF graph for Western province series

From PACF graph in Figure 3, it can be seen that the model may contain AR(1) and AR(2) terms as corresponding spikes at lag 1 and lag 2 are significant for the series.

Several models, include the combination of the above terms, are tried using EViews software. However, the models which satisfied conditions of diagnostic tests are summarised in the following table Table 2.

Table 2. Summary table of selected models for Western province

MODEL	AIC, SBC	DW	R^2	Skewness, Kurtosis	LM Test	White's general test
1	15.79, 15.90	1.97	0.42	0.74, 3.23	0.23	0.51
2	15.54, 15.72	2.05	0.58	0.06, 2.85	0.64	0.79
3	15.74, 15.91	2.10	0.49	0.73, 2.93	0.52	0.30

As per the results in the Table 2, the p values of White's general test and LM test suggest that the error terms have constant variance and no serial correlation among them.

Based on the Skewness and Kurtosis statistics, it can be confirmed that the error terms follow normality. DW values in all three models show that, the randomness of error terms.

However, second model has the highest R^2 value, lowest AIC and SBC values. Therefore the second model is selected as the best model. Accordingly, the best fitted model for the Western province is:

$$Y_t = 1290.43 + 0.96 * Y_{t-1} - 0.43 * Y_{t-2} - 0.16 * e_{t-1} + \\ 0.88 * e_{t-5} + e_t$$

Table 3. Comparison of estimated incidences with actual incidences for Western province

Month	January 2015	February 2015	March 2015	MAPE
Estimated incidence	3496.32	1671.49	686.27	14.68 %
Actual incidences	2869	1754	828	

According to the MAPE value in Table 3, it can be suggested that the fitted model is a better model for Western province as MAPE show nearly 15%.

B. Results of Colombo district

Almost the same procedure and similar arguments are followed for the series of Colombo district incidences too. The relevant results are presented as follows:

According to the records available at Epideminology unit of Ministry of Health, as average 849.67 incidences are recorded in Colombo district and which is about 27.66% of the total incidences. Hence it can be claimed that one fourth of the total incidences are reported in Colombo district. To confirm this result the following hypothesis test is performed.

$$H_0 : \rho \le 0.25 \ VS \ H_1 : \rho > 0.25$$

Test Statistic Z= 0.46<1.64 (table Z value at 5% level of significance)

Therefore it can be confirmed that one fourth of total incidences are reported from Colombo district. Thus it can be concluded with 95% confidence that one fourth of total incidences are reported in Colombo district.

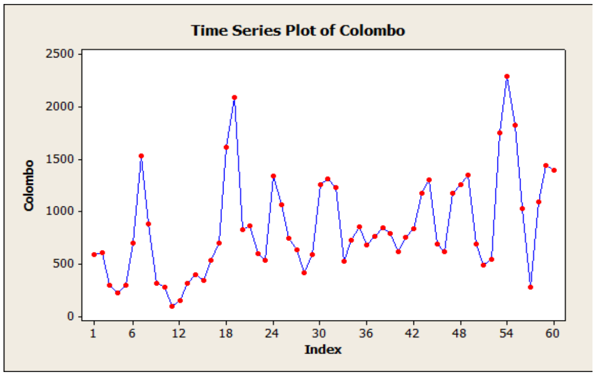

Figure 4. Time series plot of Colombo District

Plot in the Figure 4 indicates that, there is no seasonal or cyclic pattern in the Colombo district series. Here also it seems that there may be a trend. Thus ADF test is used to check whether there is a trend in Colombo district series.

Table 4. ADF Test statistics for Colombo district series

	t-Statistics
ADF test statistics	-5.386044
Test critical values: 1% level	-3.548208
5% level	-2.912631
10% level	-2.594027

By comparing t- values of the ADF test statistic in all three significance level as appear in Table 4, it can be confirmed with 99% confidence that there is no trend in the series.

The following both figures Figure 5 and Figure 6 are obtained using MINITAB to check the stationary condition of the series meanwhile to guess the terms which should be included in the ARMA models.

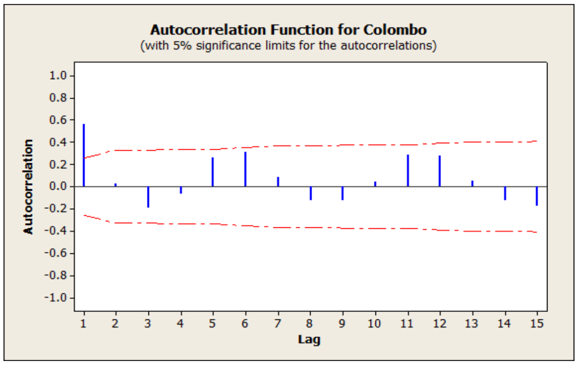

Figure 5. ACF graph for Colombo district series

From ACF graph in Figure 5, it can be seen that the model may contain MA(1) term as the corresponding spike only at lag 1 is significant in ACF graph for the Colombo district series.

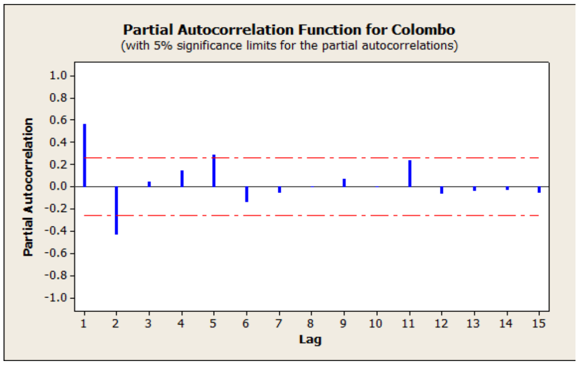

Figure 6. PACF graph for Colombo district series

From PACF graph in Figure 6, it can be observed that the model may contain AR(1), AR(2), AR(5) and AR(11)terms as corresponding spikes at lag 1 lag 2 , lag 5 and lag 11 are significant for the series of Colombo district.

Several models, include the combination of the above terms, are tried using EViews software. However, the models which satisfied conditions of diagnostic tests are only considered and summarised in the following table Table 5.

Table 5.Summary table of selected models for Colombo district

Month	January 2015	February 2015	March 2015	MAPE
Estimated incidence	1684.49	997.64	675.00	15.00 %
Actual incidences	1790	1088	516	

According to the results appear in the Table 5, from p-values of LM test and White's General test, it can be confirmed that the error in all three models have constant variance as well as which have no serial correlation.

Further skewness and kurtosis values indicate that the error terms in all three models are normally distributed.

DW values and R^2 values are almost equal in all three models. However by comparing AIC and SBC values, the first model is selected as the best model. Accordingly, the best fitted model for the Colombo district is:

$$Y_t = 896.26 + 0.8 * Y_{t-1} - 0.47 * Y_{t-2} - 0.51 * Y_{t-3} + \\ 0.4 * Y_{t-4} + 0.94 * e_{t-3} + e_t$$

Table 6.Comparison of estimated incidences with actual incidences for Colombo district

MODEL	AIC, SBC	DW	R^2	Skewness, Kurtosis	LM Tet	White's General Test
1	14.43, 14.65	2.05	0.62	0.40, 3.00	0.58	0.31
2	14.54, 14.75	1.98	0.58	0.37, 3.61	0.46	0.54
3	14.42, 14.71	2.01	0.65	0.50, 3.40	0.84	0.61

Based on the MAPEvalue in Table 6, it is suggested that the model developed for Colombo district is a better model as MAPE value gives 15%.

IV. CONCLUSION

Since the Western part of Sri Lanka is mostly affected by the dengue, more attention is needed for this part of the country to take necessary actions to control the incidences. Further in order to provide treatment for those who will be infected by dengue fever, an appropriate mechanism to forecast the incidences is needed.

Therefore, the model for the Western province of Sri Lanka, which is the mostly affected province in the country, is ARMA(2, 5) and fitted model is:

$$\hat{Y}_t = 1290.43 + 0.96 * Y_{t-1} - 0.43 * Y_{t-2} - 0.16 * e_{t-1} + \\ 0.88 * e_{t-5} \qquad (\text{MAPE} = 14.68\%)$$

Similarly, the model for Colombo district of Sri Lanka, which is the mostly affected district in the country, is ARMA(4, 3) and fitted model is:

$$\hat{Y}_t = 896.26 + 0.8 * Y_{t-1} - 0.47 * Y_{t-2} - 0.51 * Y_{t-3} + \\ 0.4 * Y_{t-4} + 0.94 * e_{t-3} \qquad (\text{MAPE} = 15.00\%)$$

By using these two ARMA models, one can forecast the dengue incidences in Western province as well as in Colombo district for the near future and can take actions accordingly.

V. REFERENCES

Arul E, Say BT, Annelies WS, and David M (2012), Comparing Statistical Models to Predict Dengue Fever Notifications, *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 758674, 6 pages, 2012. doi:10.1155/2012/758674

Centre for Dengue Research (2012),*University of Sri Jayawardenapura, Sri Lanka* [online] Available from http://www.cdrsrilanka.com/ [Accessed: 15th June 2015]

Distribution of notification dengue cases by month (2015), *Epidemiology Unit, Ministry of Health, Sri Lanka* [online] Available from http://www.epid.gov.lk/web/index.php?option=com_casesanddeaths&Itemid=448&lang=en [Accessed: 15th June 2015]

Goto K, Kumarendran B, Mettananda S, Gunasekara D, Fujii Y, Kaneko S (2013) 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

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

Murugananthan K, Kandasamy M, Rajeshkannan N, Noordeen F (2014), Demographic and clinical features of suspected dengue and dengue haemorrhagic fever in the Northern Province of Sri Lanka, a region afflicted by an internal conflict for more than 30 years—a retrospective analysis, *International Journal of Infectious Diseases*,Vol 27, 32-36.

National Plan of Action Prevention and Control of Dengue Fever 2005-2009 (2010), *Epidemiology Unit, Ministry of Health Sri Lanka* [online] Available http://www.epid.gov.lk/web/images/pdf/Circulars/latest_draft_poa_for_dfdhf.pdf [Accessed: 15th June 2015]

Sirisena PDNN, Noordeen F, (2014), Evolution of dengue in Sri Lanka- changes in the virus, vector, and climate, *International Journal of Infectious Diseases*,Vol 19, 6-12.

Special Second Mosquito Control Programme- Western province and selected districts 2015 Phase II- Summary (2015), *National Dengue Control Unit, Ministry of Health Sri Lanka* [online] Available http://www.dengue.health.gov.lk/ [Accessed: 15th June 2015]

Weekly Epidemiological Report (2009), Current Situation and Epidemiology of Dengue in Sri Lanka, *Epidemiology Unit, Ministry of Health Sri Lanka* [online] Available http://www.healthedu.gov.lk/web/images/pdf/msp/current_situation_epidemiology_of_dengue.pdf [Accessed: 15th June 2015]

BIOGRAPHY OF AUTHORS

Mr. S. R. Gnanapragasam attached to the Department of Mathematics and Computer Science, The Open University of Sri Lanka as a Lecturer (Probationary). Graduated (B. Sc Special in Mathematics) from University of Peradeniya in 2005 and received a Master degree in Applied Statistics from the same University in 2012. He has nearly 10 years of experience at University level teaching for undergraduate level as Temporary Lecturer and Lecturer (Probationary). He has some publications in the field of Applied Statistics.

Mr. T. M. J. A. Cooray is serving as a Senior Lecturer in the Department of Mathematics, University of Moratuwa. He has obtained a M. Phil (Moratuwa) in 2003 after completion of a M. Sc (Colombo) in 2000 and Postgraduate Diploma in Mathematics (Peradeniya) in 1979. He graduated from University of Peradeniya with B. Sc in 1978. He has 28 years of experience at University of Moratuwa as an Assistant Lecturer, Lecturer and Senior Lecturer in undergraduate and post graduate levels. He has widely published his research work in his specialized area of Time Series and Operational Research.