JSc EUSL (2020) Vol. 11 No. 1, p 11-22 *DOI: http://doi.org/10.4038/jsc.v11i1.24*

Print- ISSN: 1391-586X, E-ISSN: 2602-9030, Copyright © 2020 by Faculty of Science, Eastern University, Sri Lanka.

EXPONENTIAL SMOOTHING ON FORECASTING DENGUE CASES IN COLOMBO, SRI LANKA

Attanayake, A.M.C.H [1*]., Perera, S.S.N.,[2] and Liyanage, U.P[3]

[1,3]Department of Statistics & Computer Science, Faculty of Science, University of Kelaniya.

[2] Research & Development Centre for Mathematical Modelling, Faculty of Science, University of Colombo.

ABSTRACT

Prediction of number of people to be infected is an essential component in studying any leading diseases. Particularly it is important in dengue disease as it is the most critical mosquito-borne viral disease in the world. The number of reported dengue cases gradually increased all over the world as well as in Sri Lanka. In Sri Lanka, the majority of dengue cases reported in the Colombo district. The authors applied exponential smoothing technique in order to model and forecast dengue cases in Colombo, Sri Lanka. Data consist of monthly reported dengue cases in Colombo district from January 2010 to May 2019. January 2010 to February 2019 data used for model building and rest of the data used for model validation. Both original cases and log transformed cases considered for modelling and Holt Winters smoothing suits well with both cases. Best model in each case and finally the most parsimonious model within these two best models were selected by considering AIC, BIC, MAE, RMSE and MAPE measures. The most parsimonious model fits on log transformed dengue cases. Using the most parsimonious model predictions were made for June to August 2019. It can be concluded that the best model able to fit on the data in an adequate level and reported dengue will increase slowly during the prediction period.

.

Key Words: Dengue, Exponential Smoothing, Prediction

**Corresponding author: succ@kln.ac.lk*

http://orcid.org/0000-0002-5200-3751

1.0 INTRODUCTION

Dengue is one of the fastest spreading infectious diseases around the world. It is caused by the bites of infected mosquitos which is identified as Aedes aegypti mosquitos. Infected people with dengue fever are drastically increased over recent years in all over the world. World Health Organization (WHO) estimates that 390 million people world-wide getting infected annually with dengue [1]. The most affected areas are Southeast Asia, the Americas and Western Pacific. From the annual estimates of the dengue disease, nearly 500, 000 cases transformed in to more severe form of the dengue; named as dengue hemorrhagic fever and resulted 25,000 deaths annually worldwide. Therefore, dengue controlling and management actions are necessary to reduce the burden of the dengue disease.

Sri Lanka is an island located in the Indian Ocean. Dengue found in Sri Lanka in 1960 [2] and gradually increased the number of infected people over years. Approximately 43 % of the dengue fever cases were reported from the Western Province of Sri Lanka in 2017 and the usually the most affected area with the highest number of reported dengue cases is the Colombo district [3]. The other highly affected areas are Galle, Jaffna, Kaluthara and Gampaha. 51,659 number of dengue cases reported in 2018 and 36 858 suspected dengue cases were reported to the Epidemiology Unit of Sri Lanka from all over the island in March 2019 [3].

The modelling and predicting the number of reported dengue cases will be useful to understand the dynamics of the disease and thus to control it. Exponential Smoothing technique is one of the powerful forecasting techniques available in the area of univariate time series analysis. The invention of exponential smoothing technique is in late 1950s which is during the study of Robert Goodell Brown [4] and expanded the technique by Charles C. Holt [5]. As the name implies the technique uses exponentially weighted observations in order to make predictions. More recent the observation promotes to get a higher weight than older observations. Hence, this exponential smoothing technique is useful in short-term forecasting. Several methods are available within the exponential smoothing technique such as simple exponential smoothing, double exponential smoothing, Holt-Winters additive and multiplicative methods, etc. These methods are widely applying in almost all fields because of its simplicity, accuracy and it also assumes minimum assumptions. Particularly, many applications of exponential smoothing techniques in the field of epidemiology can be found in literature [6, 7, 8]. But there are limited numbers of studies in the literature on application of exponential smoothing on dengue fever specially in Sri Lankan context.

In this study, monthly dengue cases were predicted using exponential smoothing for Colombo district. Both original data set and log transformed data set considered for modelling under exponential smoothing with the purpose of finding the best model to predict the dengue disease in Colombo. Multiple model selection criterions were considered in order to select the best prediction model. Availability of an effective prediction model will helpful in anticipating the dengue and to make timely actions on controlling the dengue incidence.

2.0 MATERIALS AND METHODS

2.1 Secondary Data and Model Selection

Monthly reported dengue cases in Colombo district were acquired from the Epidemiology Unit of Ministry of Health, Sri Lanka from January 2010 to May 2019. Exponential smoothing models were fitted for both monthly reported dengue cases as well as for the log-transformed monthly reported dengue cases by considering data from January 2010 to February 2019. Data from March to May in 2019 were used for model validation. Forecasted values for three months from June to August in 2019 were generated from the best exponential smoothing model.

2.2 Statistical Tests and Methods

The following tests and methods were used in the study:

2.2.1 Time Series Plot: A graph of values against time and usually uses to extract meaningful characteristics of data. This graph can be used to detect outlying observations, patterns of data such as trends, seasonal and cyclic variations.

2.2.2 Augmented Dickey Fuller (ADF) Test: A test uses to detect whether a series has a unit root or not. The test statistics for the model $Y_t = \rho Y_{t-1} + u_t$ is $\frac{\hat{\rho}}{SE(\hat{\rho})} \sim t_{n-1}$ where $-1 < \rho < 1, u_t$ the white noise is, *n* is the number of observations. Null hypothesis is series is non-stationary.

2.2.3 Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test: Use for testing a null hypothesis that an observable time series is stationary around a deterministic trend against the alternative of a unit root. The KPSS test is based on a linear regression. It breaks up a series into three parts: a deterministic trend (β_t), a random walk (r_t), and a stationary error (ε_t), with the regression equation:$x_t = r_t + \beta_t + \varepsilon_t$.
If the data is stationary, it will have a fixed element for an interceptor the series will be stationary around a fixed level.

2.2.4 Autocorrelation Function (ACF)
The coefficient of correlation between two values in a time series is called the autocorrelation function. For example, the ACF for a time series x_t is given by:

$$Corr(x_t, x_{t-k}), k = 1,2, \ldots$$

This value of k is the time gap being considered and is called the lag. A lag 1 autocorrelation ($k = 1$ in the above) is the correlation between values that are one time period apart. More generally, a lag k autocorrelation is the correlation between values that are k time periods apart.

2.2.5 Partial Autocorrelation Function (PACF)
The partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the ACF, which does not control for other lags.

2.2.6 Box-Pierce test

A statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags.

H_0: The data are independently distributed

H_1: The data are not independently distributed; exhibit serial correlation.

The test statistics is $Q = n\sum_{k=1}^{h}\rho_k^2$ where ρ_k is the sample autocorrelation at lag *k*, and *h* is the number of lags being tested. Under H_0**,** the statistic Q follows χ_h^2**.** For significance level α, the critical region for rejection of the hypothesis of randomness is: $Q > \chi_{1-\alpha,h}^2$ where *h* is the degrees of freedom.

2.2.7 Exponential Smoothing

In exponential smoothing recent observations are weighted more heavily than older observations.

2.2.7.1 Simple Exponential Smoothing (SES)

The simplest of the exponentially smoothing methods is called simple exponential smoothing (SES). The SES is suitable for modelling if the data do not represent a trend or seasonal pattern. The weight of each observation is determined by using a smoothing parameter, α. For a data set of t number of observations where the last observation is y_t and the predicted value at time *t+1*; y_{t+1} is determined as follows:

$$y_{t+1} = \alpha y_t + \alpha(1-\alpha)y_t + \cdots + \alpha(1-\alpha)^{t-1}y_1,$$

where $0 < \alpha \le 1$. In the *component form* of this model the following set of equations can be used:

$$y_{t+1} = l_t$$
$$l_t = \alpha y_t + (1-\alpha)l_{t-1}$$

2.2.7.2 Double Exponential Smoothing

This method is more applicable when the data represent a trend. In addition to the smoothing parameter used in SES there is another parameter β to capture the trend. The k step-ahead forecast is generated by concatenating the level estimate at time *t* as L_t and the trend estimate (which is assumed additive) at time *t* as T_t as follows:

$$y_{t+1} = L_t + kT_t,$$

where the level estimate L_t and trend estimate T_t will update by using updating equations with two smoothing parameters α and β as given below:

$$L_t = \alpha y_t + \alpha(1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)(T_{t-1})$$

The first equation represents the level at time *t* is a weighted average of the actual value at time *t* and the level in the earlier period, adjusted for trend. The second equation represents the trend at time *t* is a weighted average of trend in the earlier period and the more recent information. The parameters *α* and *β* are in between *0* and *1*.

To capture a multiplicative trend, the following changes must be made to the above equations:

$$y_{t+1} = L_t * kT_t$$
$$L_t = \alpha y_t + \alpha(1-\alpha)(L_{t-1} * T_{t-1})$$
$$T_t = \beta(L_t/L_{t-1}) + (1-\beta)(T_{t-1})$$

2.2.7.3 Holt-Winters Seasonal Smoothing

If a data set comprises with both trend and seasonality this smoothing will useful in making predictions. This method has two options to capture additive seasonality or multiplicative seasonality. For an additive Holt-Winters model a general equation is:

$$y_{t+1} = L_t + kT_t + S_{t+k-m}$$

Three smoothing parameters; the level, trend and season will update using the following updating equations:

$$L_t = \alpha(y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)(T_{t-1})$$
$$S_t = \gamma(y_t - L_t) + (1-\gamma)(S_{t-m})$$

Where α, β and γ are the three smoothing parameters to capture pattern, trend and seasonality respectively. All three parameters are in between *0* and *1*.

For multiplicative seasonality the following equations can be used:

$$y_{t+1} = (L_t + kT_t)S_{t+k-m}$$
$$L_t = \alpha(\frac{y_t}{S_{t-m}}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)(T_{t-1})$$
$$S_t = \gamma(y_t/L_t) + (1-\gamma)(S_{t-m})$$

By considering trend, seasonality and resulting error structure as either additive or multiplicative, varies models can be constructed and validated.

2.2.8 Model Selection Criteria

When more than one model fit on the data in an adequate level some model selection criteria such as; Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) can be used to select the most parsimonious model

2.2.8.1 Akaike Information Criteria (AIC)

AIC is calculated by using the formula:

$$AIC = -2\ ln(L) + 2k,$$

where *k* is the number of parameters in the model and *L* be the maximum value of the likelihood function for the model. A model with the minimum AIC value will be the best model.

2.2.8.2 Bayesian information criterion (BIC)

BIC is one of the model selection criterion to select a model among a finite set of available models. A model with the minimum BIC will be the best model. The formula for BIC is given below:

$$BIC = -2\, ln(L) + ln(n)\, k,$$

where *L* be the maximum value of the likelihood function for the model, *n* is the sample size and *k* is the number of parameters in the model.

2.2.8.3 Root Mean Square Error (RMSE)

RMSE is an accuracy measure which represents the difference between two data sets; predicted values and actual values. That represents error or residual.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} = \sqrt{Mean\ Squared\ Error},$$

where e_t is the residual at time *t* and *n* is the total number of the time periods.

2.2.8.4 Mean Absolute Error (MAE)

It is the average absolute difference between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|,$$

where e_t is the residual at time *t* and *n* is the total number of the time periods.

2.2.8.5 Mean Absolute Percentage Error (MAPE)

MAPE is a popular measure of prediction accuracy which is given by the following formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{|e_t|}{y_t} * 100\%,$$

Where e_t is the residual at time *t* , y_t is the actual value at time *t* and *n* is the total number of the time periods. If calculated value of MAPE is less than 10 %, it is interpreted as excellent accurate forecasting, between 10–20 % good forecasting, between 20–50 % acceptable forecasting and over 50 % inaccurate forecasting [9].

3.0 RESULTS AND DISCUSSION

R software is mainly used for the data analysis [10]. Figure 1 displays the time series plot of monthly reported dengue cases whereas the associated Autocorrelation Function describes in Figure 2 with the Partial Autocorrelation Function in Figure 3.

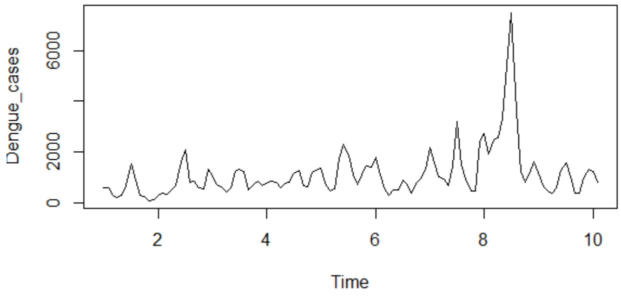

Figure 1: Time series plot of monthly reported dengue cases in Colombo, Sri Lanka

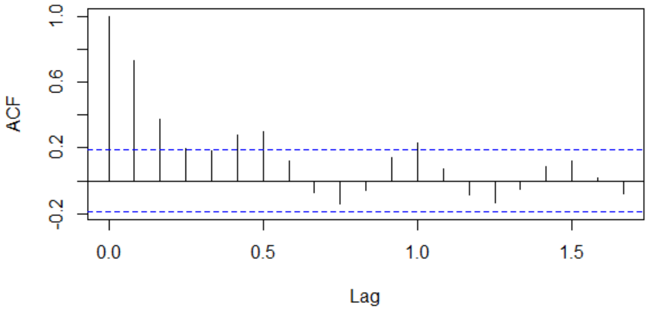

Figure 2: ACF of Monthly reported dengue cases in Colombo, Sri Lanka

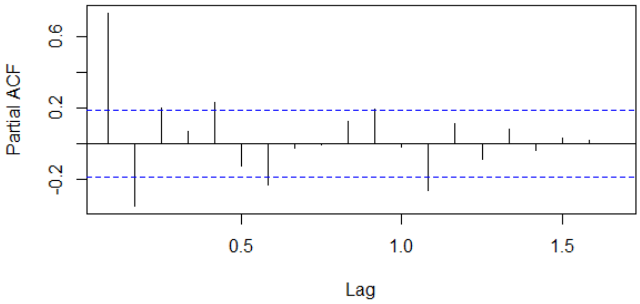

Figure 3: PACF of Monthly reported dengue cases in Colombo, Sri Lanka

According to Figure 1 reported dengue cases in the Colombo district varies in between the minimum value of 97 and maximum value of 3620 other than the highest numbers of dengue cases were reported in June and July of 2017. Both ADF and KPSS tests confirms the non-stationarity of the original dengue series at 5% significance level whereas it can be already seen from ACF and PACF plots that there is a seasonality in the series. The Box-Pierce test indicated the dependency of autocorrelations. Therefore, both simple and double exponential smoothing techniques will not be applicable for modeling the series. Hence, Holt Winters smoothing technique was applied to predict future dengue cases. All possible combinations that can be considered for modelling by changing multiplicative and additive structures for all error, trend and seasonality of the series were implemented. The optimal model with minimum AIC, BIC, MAE, MAPE and RMSE selected as the best model to forecast future dengue cases in Colombo district. It consists with multiplicative error, multiplicative seasonality and additive structure for error. Residual analysis of that model of monthly reported dengue cases is given in Figure 4.

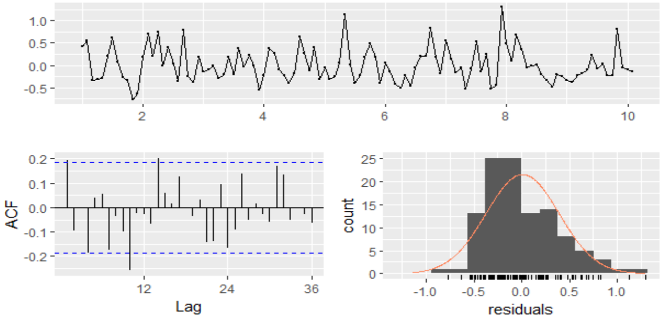

Figure 4: Residual analysis of the model selected for monthly reported dengue cases

To minimize the variability present in the original series which is represented in Figure 1, log transformed monthly reported dengue cases also considered for modelling using exponential smoothing. When examining the time series plot (Figure 5), ACF (Figure 6) and PACF (Figure 7) of the log transformed monthly reported dengue cases it can be seen that the transformed series also represent non-stationarity. Further it is confirmed by ADF and KPSS tests at 5% level of significance. The Box-Pierce test indicated the dependency of autocorrelations. Because there is a seasonality in the transformed series and it is also non stationary, Holt Winters smoothing may be more appropriate for modelling.

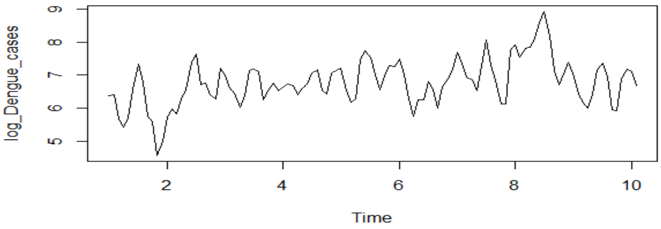

Figure 5: The time series plot of log-transformed monthly reported dengue cases in Colombo, Sri Lanka

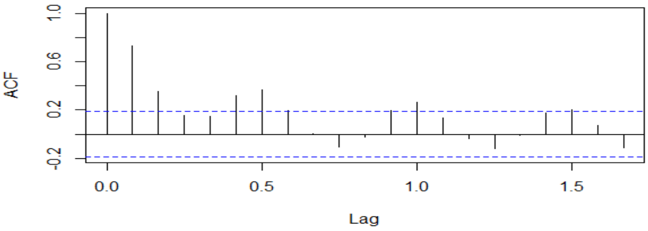

Figure 6: ACF of Log-Transformed Monthly reported dengue cases in Colombo, Sri Lanka

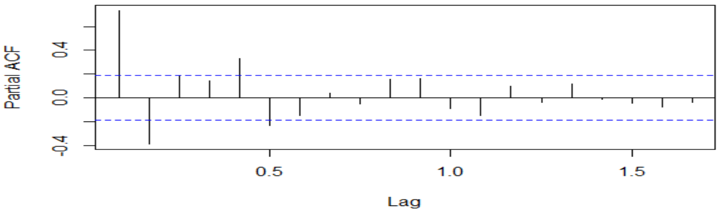

Figure 7: PACF of Log-Transformed Monthly reported dengue cases in Colombo, Sri Lanka

By changing the additive or multiplicative structure for trend, seasonality and error, all the possible combinations for modelling were considered and optimal model with minimum AIC, BIC, MAE, RMSE and MAPE selected as the best model to predict the monthly reported dengue cases in Colombo, Sri Lanka. The best model for the log transformed series was with additive error, additive seasonality with no trend. Although there is no strong assumption of normality and independency of errors under exponential smoothing, the residual analysis of the best model for transformed series is given in Figure 8. Normality of residuals confirms in Figure 8 than in Figure 4.

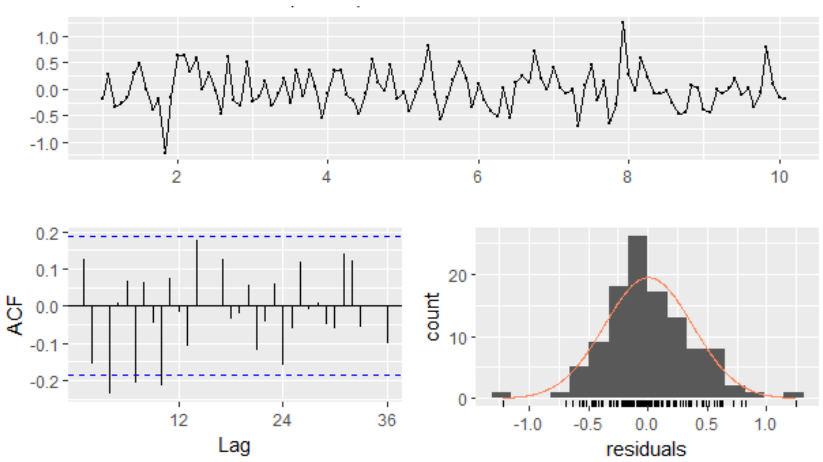

Figure 8: Residual analysis of the model selected for transformed monthly reported dengue cases

Among the best models selected under monthly cases and transformed monthly cases the most parsimonious model to predict monthly reported dengue cases was selected by considering accuracy measures which as summarized in Table 1. In both cases using two models predicted values were generated for March to August in 2019 and for March to May 2019 data were used to validate (test) the models.

By comparing the summary measures in Table 1 it can be conclude that minimum AIC, BIC, RMSE, MAPE and MAE values given by the model fitted on log-transformed monthly reported dengue series.

Table 1: Summary measures of the best models selected under monthly and transformed monthly dengue cases

	Monthly reported dengue cases		**Transformed monthly reported dengue cases**	
Smoothing parameters	α = 0.6362 β = 1*e-04 γ = 1*e-04		α = 0.817 γ = 1*e-04	
	Training set	Test set	Training set	Test set
RMSE	408.4125	102.1103	0.3688	0.2089
MAPE	34.9930	12.2949	4.2701	2.5793
MAE	297.0214	95.7012	0.2819	0.1718
AIC	1843.719		327.6161	
BIC	1889.627		368.1233	

It further minimizes the prediction errors on the test set. Therefore, the best model to forecast monthly reported dengue cases in Colombo district is Holt Winters exponential smoothing model fitted on log transformed dengue series. The forecasted values for March to August 2019 given by the best model is shown in Figure 9 with 80 % and 95 % confidence intervals.

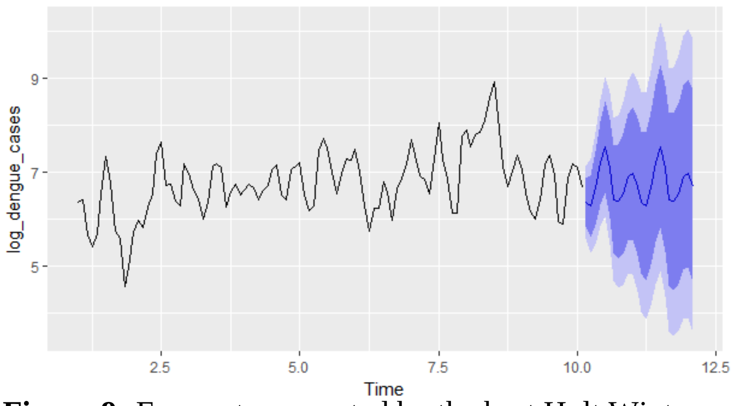

Figure 9: Forecasts generated by the best Holt Winters model

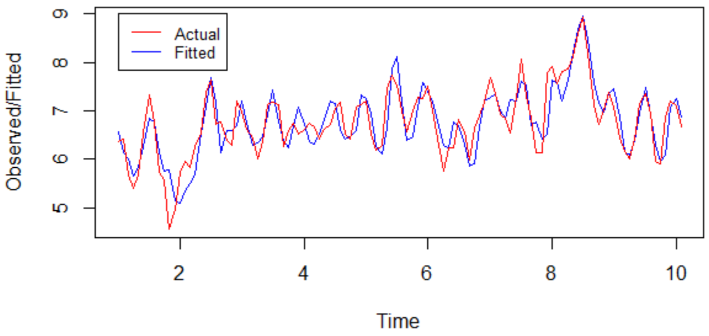

Figure 10: Actual log transformed Dengue cases and fitted values generated by the best model

Several model selection criterions were considered in this study in order to select the best model to forecast dengue cases rather than considering one or two criterions. The forecasted values were generated by the best model for the period of March to August 2019. Since the predictions cover the upcoming periods in 2019 it will be more useful in controlling the disease and managing the resources related to the dengue disease rather than some models fit in other researches use past data that do not cover the current year 2019. By considering the forested values of the model it can be concluded that the monthly dengue cases to be reported in the upcoming months (June to August in 2019) will increase slowly in Colombo district.

4.0 CONCLUSION

This study successfully models the monthly reported dengue cases in Colombo, Sri Lanka through exponential smoothing technique with the aim of forecasting future dengue cases. Specially, Holt Winters smoothing technique suits well with the available data under the area of exponential smoothing. Both original data and log transformed data considered for modelling with the purpose of finding the best model to predict the dengue disease in Colombo. The best model to predict monthly dengue cases in Colombo district is the Holt Winters exponential smoothing model fitted on log transformed series which exists additive error, additive seasonality with no trend. The forecasted values generated by the best model for the March to August 2019 are 580, 536, 784, 1342, 1898 and 1192. The forecasted values may be useful in taking actions towards controlling the dengue cases in Colombo, Sri Lanka.

5.0 ACKNOWLEDGEMENT

This research was supported by the grant number RPHS/2016/D05 of National Science Foundation (NSF), Sri Lanka.

6.0 REFERENCES

[1] World Health Organization (2019). Dengue and Severe Dengue, Retrieved from http://www.who.int/mediacentre/factsheets/fs117/en/.

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

[3] Epidemiology Unit, Ministry of Healthcare and Nutrition, Sri Lanka (2019). Dengue Update, Retrieved from http://www.epid.gov.lk.

[4] Brown, Robert G. (1956). Exponential Smoothing for Predicting Demand, Cambridge, Massachusetts: Arthur D. Little Inc, 2.

[5] Holt, Charles, C. (1957). Forecasting Trends and Seasonal by Exponentially Weighted Averages, Office of Naval Research Memorandum, Carnegie Institute of Technology. Reprinted in Holt, Charles C.,(2004), Forecasting Trends and Seasonal by Exponentially Weighted Averages, *International Journal of Forecasting*, **20(1):**5–10.

[6] Payam, A., Ali G., Elaheh, Z., Majid, S., Mohammad, E., Zahra, and T., Saeid, Y. (2018). Modelling the Frequency of Depression using Holt-Winters Exponential Smoothing Method, *Journal of Clinical and Diagnostic Research*, **12: 10**.

[7] Zhang, X., Zhang, T., Young, A., A., and Li, X. (2014). Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data, *PLoS ONE*, **9:2**. https://doi.org/10.1371/journal.pone.0091629

[8] Ghaffari, M.E., Ghaleiha, A., Taslimi, Z., Sarvi, F., Amini, P., and Sadeghifar, M. (2017). Forecasting schizophrenia incidence frequencies using time series approach, *Int Clin Neurosci J*, **4(4):** 152-156. DOI: 10.15171/icnj.2017.06.

[9] Lewis, C.D., (1982). Industrial and Business Forecasting Methods, London, Butterworth, **2(2)**:194-196.

[10] R Core Team, (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **7(5)**.

[11] Eva O., and Oskar O. (2012). Forecasting Using Simple Exponential Smoothing Method, *Acta Electrotechnical et Informatica*, **12(3):** 62–66. DOI: 10.2478/v10198-012-0034-2.