

Multiple Linear Regression Models on Interval-valued Dengue Data with Interval-valued Climatic Variables

A.M.C.H. Attanayake¹, S.S.N. Perera² and U.P. Liyanage³

¹Department of Statistics & Computer Science, Faculty of Science,
University of Kelaniya, Sri Lanka;
Email : succ@kln.ac.lk

²Research & Development Centre for Mathematical Modelling, Department of Mathematics,
Faculty of Science, University of Colombo, Sri Lanka;
Email : ssnp@maths.cmb.ac.lk

³Department of Statistics & Computer Science, Faculty of Science,
University of Kelaniya, Sri Lanka;
Email : liyanage@kln.ac.lk

ABSTRACT

Reported dengue fever cases are increasing day by day in the world as well as in Sri Lanka. Model, Prediction and Control are three major parts of the process of analysis of the dengue incidence which leads to reduce the burden of the dengue. There is an increasing trend in the applications and developments in interval-valued data analysis over recent years. Particularly, under regressions there have been developed various techniques to handle interval-valued dependent and independent variables. Representation of data as intervals is very much useful to capture uncertainty and missing details associated with variables. Further, the predictions in intervals suit well when the situations of exact forecasts may not be necessary. In this study interval-valued dengue data with interval-valued minimum temperature, maximum temperature and rainfall from 2009 to 2015 in the Colombo district, Sri Lanka were modelled using three interval valued regression procedures, namely, Center Method (CM), Center and Range Method (CRM) and Constrained Center and Range Method (CCRM). Predicted dengue cases in a range is particularly important because actions taken towards controlling the dengue do not depend on exact number but on magnitude of the values represented in the interval. Data in the year 2016 used for the validation of the models which is developed under three methods. Root of the mean square error, coefficient of determination as well as square root of variance of the models were used to select the best procedure to predict dengue cases. Among the three regression procedures both CRM and CCRM perform well in predicting monthly dengue cases in Colombo.

Keywords: Dengue, Interval-valued data, Regression.

Mathematics Subject Classification: 62J12, 62G99

1. INTRODUCTION

Dengue is one of the life-threatening diseases in the entire world which records large number of mortalities. The disease spreads through mosquitos and higher the mosquito density increase the burden of the disease. Several causing types of the virus of the disease found up to date but one person can infect with all of the types of the disease during his/her lifetime [1]. This fact shows the severity and complexity of the dengue virus.

World Health Organization (WHO) found that 390 million people in the world suffering with the dengue disease in each year [1]. Further, according to their estimates more than 0.4% of the population in the world, including more than 100 countries are at higher risk of dengue fever. These estimates should be higher than the recorded values because there be many misclassified and not reported dengue cases left in the underground. There is an increase of the number of suffered and reported dengue counts in almost all of the dengue prone counties in the world.

In Sri Lanka, the spread of the dengue increased year by year. It has obtained the attention of the public domain due to its heavy spreading nature and as well as its severity. Within the entire country, generally the highest number of dengue counts found in the Western province whereas the most affected district recorded as the Colombo [2]. Sri Lanka has experienced a marked increase in reported dengue cases over recent years.

Many studies in the past years stated the correlation between climatic variables with the dengue incidence [3, 4, 5, 6]. Temperature and rainfall are closely related to increase the reported dengue cases. The stages of the life cycle of the dengue mosquito shorten with higher temperature levels [7]. This fact prone to increase the mosquito population and hence to increase the reported dengue spread among human population. Rainfalls provide suitable breeding sites for mosquitos resulted high density of mosquitos which leads to increase the infected human population with dengue.

Since dengue is a life threatening, extensively spreading disease and there is no exact treatment for the disease, it is important to understand and propose controlling strategies for the dengue. Regression analysis is one of the powerful areas that can be applied to model and predict dengue incidence. Under the area of regression analysis, multiple linear regression is a statistical technique use to model the connection of an output variable with several input variables. The coefficients of the regression model parameters usually estimate by using ordinary least squares method which minimizes the error represent by the actual and fitted values of the output variable. By considering reported dengue counts as the output variable and correlated climatic variables as input variables will provide platform for effective dengue modeling through multiple regression.

Common practice is to apply regressions on data of point values [8,9]. Interval-valued regression comes to play during the last decade. Interval-valued datum represents by a lower bound and an upper bound for an interval rather than representing it as an exact or a point value. These kinds of representations provide additional information for the spread of the value that can be used to make more precise decisions. Interval of values may contain uncertainties as well as observational and rounding errors associated with a point value. Analysis of intervals of values leads to make better predictions which will provide approximations for a point value. Many real life scenarios compromise with uncertainties. Therefore, it is unrealistic to represent such uncertainties as point values. Hence, in such cases It is useful to estimate values as intervals.

The first linear regression model called Center Method (CM) for interval data prediction was found by Billard and Diday in the year 2000 [10]. In CM there is no assurance of lower level of predicted value is minor than the upper level. Therefore, Lima Neto and De Carvalho in 2008 [11] formulated the center and range method (CRM) to capture the additional information provided by the range of the intervals.

CRM formulates two separate predictions as prediction of center and prediction of range which is finally aggregate to find lower prediction level and upper prediction level of the output variable. Lima Neto and De Carvalho in 2010 [12] invented the constrained center and range method (CCRM) by introducing a non-negative constraint on the coefficients of regression model which is developed on range values. All these mentioned methods are based on ordinary least squares regression and recently many methods and applications have been developed on interval-valued regressions [17,18].

Modelling and predicting dengue cases are highly depending on climatic factors which are associated with uncertainties and many unknown parameters which cannot be estimated may influence the spread of the disease. In such a scenario interval valued representation of dengue and climatic factors are very much useful. However, interval valued multiple regression is not yet utilized to predict dengue cases, especially in Sri Lankan context. This study is an attempt to analysis and compare the feasibility of interval valued multiple regression techniques to predict dengue cases.

In this study, interval valued maximum temperature, minimum temperature and rainfall are considered as independent whereas interval valued dengue cases considered as dependent. CM, CRM and CCRM techniques were used to predict the dengue cases and comparison were done in-order to identify significance of each technique, hence to determine the best technique. The section 2 summarizes the theories and methods behind the analysis whereas the section 3 describes the results and interpretations found through the analysis which follows the final conclusion of the study.

2. METHODS AND TECHNIQUES

2.1 Secondary Data

Monthly lower limit and upper limit for the dengue count in Colombo, Sri Lanka were found from the data acquired from the Epidemiology Unit, Ministry of Health, Sri Lanka for the duration of January 2009 to December 2016. Monthly lower limit and upper limit for the rainfalls, monthly lower limit and upper limit for the maximum temperature and monthly lower limit and upper limit for the minimum temperature in Colombo were further identified from the collected data of Department of Meteorology, Sri Lanka for the same period as January 2009 - December 2016.

Multiple linear regression models using CM, CRM and CCRM were fitted by considering data representation: January 2009 - December 2015. January - December data in 2016 were used for validation of models.

2.2 Statistical Tests and Methods

The following tests and methods were used in the study:

2.2.1 Multiple Linear Regression

The method of Multiple Linear Regression can be applied to find the connection of an output variable with multiple input variables [13]. The multiple linear regression model equation of p number of input variables with the output variable is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} \quad \text{for } i = 1, 2, \dots, n$$

y_i - i^{th} value of the output variable, x_{ij} - j^{th} datum of i^{th} input variable and β_0 through β_p are regression parameters which usually find using least squares estimation.

2.2.2 Method of Least Squares (MLS)

The MLS is use to find the regression coefficients by minimizing the squared differences of observed figures and the fitted figures. It is more convenient to use matrix notations to apply MLS.

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \beta = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_p)$$

n - number of data points and p - count of input variables.

The MLS estimator for β is:

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

The corresponding regression line is given by $Y = X\beta$.

2.2.3 Center Method (CM)

The Center Method was introduced by Bilard and Diday (10). Let Y and X are two variables in interval representation, related as in the following relationship:

$$\begin{aligned} Y^c &= X^c \beta + \varepsilon^c \\ Y^c &= (Y_1^c, \dots, Y_n^c), \\ X^c &= ((x_1^c)', \dots, (x_n^c)'), \\ (x_i^c)' &= (1, x_{i1}^c, \dots, x_{ip}^c) \\ \beta &= (\beta_0, \dots, \beta_p)' \\ \varepsilon^c &= (\varepsilon_1^c, \dots, \varepsilon_n^c) \\ x_{ij}^c &= \frac{a_{ij} + b_{ij}}{2} \\ Y_i^c &= (Y_L + Y_U)/2 \end{aligned}$$

Y_L - lower limit of the interval

Y_U - upper limit of the interval

The MLS estimator of β is:

$$\hat{\beta} = (X^c X^c)^{-1} (X^c Y^c)$$

The upper predicted and lower predicted values are:

$$\hat{Y}_L = X_L' \hat{\beta}$$

$$\hat{Y}_U = X_U' \hat{\beta}$$

2.2.4 Center and Range Method (CRM)

L. Neto with D. Carvalho invented the CRM [11]. This method considers both the information of centers and ranges (spread) of the variables. Let Y and X are two variables in interval representations and related over the centers, then the relationship can be written as:

$$Y^c = X^c \beta + \varepsilon^c$$

$$Y^c = (Y_1^c, \dots, Y_n^c),$$

$$X^c = ((x_1^c)', \dots, (x_n^c)'),$$

$$(x_i^c)' = (1, x_{i1}^c, \dots, x_{ip}^c)$$

$$\beta = (\beta_0, \dots, \beta_p)'$$

$$\varepsilon^c = (\varepsilon_1^c, \dots, \varepsilon_n^c)$$

$$x_{ij}^c = \frac{a_{ij} + b_{ij}}{2}$$

$$Y_i^c = (Y_L + Y_U)/2$$

$$\hat{\beta} = (X^c X^c)^{-1} (X^c Y^c)$$

If Y and X are related over the rangers, then the relationship can be written as:

$$Y^r = X^r \beta^r + \varepsilon^r$$

$$Y^r = (Y_1^r, \dots, Y_n^r),$$

$$X^r = ((x_1^r)', \dots, (x_n^r)'),$$

$$(x_i^r)' = (1, x_{i1}^r, \dots, x_{ip}^r)$$

$$\beta^r = (\beta_0^r, \dots, \beta_p^r)'$$

$$\varepsilon^r = (\varepsilon_1^r, \dots, \varepsilon_n^r)$$

$$x_{ij}^r = \frac{b_{ij} - a_{ij}}{2}$$

$$Y_i^r = (Y_U - Y_L)/2$$

The MLS estimator of β^r is:

$$\hat{\beta}^r = (X^r X^r)^{-1} (X^r Y^r)$$

The upper predicted and lower predicted values:

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r$$

$$\hat{Y}_U = \hat{Y}^c + \hat{Y}^r$$

2.2.5 Constrained Center and Range Method (CCRM)

L. Neto with Carvalho introduced the CCRM [12]. In addition to CRM there is an additional constraint to ensure the condition $\hat{Y}_L \leq \hat{Y}_U$. The relationships over the centers and rangers as follows:

$$Y^c = X^c \beta + \varepsilon^c$$

$$Y^r = X^r \beta^r + \varepsilon^r$$

with constraint $\beta_j^r \geq 0; j = 0, 1, \dots, p$

The MLS estimators are:

$$\hat{\beta}^r = (X^{r'} X^r)^{-1} (X^{r'} Y^r)$$

$$\hat{\beta} = (X^{c'} X^c)^{-1} (X^{c'} Y^c)$$

2.2.6 Coefficient of Determination (CD)

The CD is the fraction of the variance in the output variable that can predict from one or more input variables.

The CD in the CM is:

$$CD = R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i^c - \bar{Y}^c)}{\sum_{i=1}^n (Y_i^c - \bar{Y}^c)}$$

$$\hat{Y}_i^c = (\hat{Y}_L + \hat{Y}_U)/2$$

$$CD = \bar{Y}^c = (\bar{Y}_L + \bar{Y}_U)/2$$

The CD in the CRM and CCRM is:

$$R_1^2 = \frac{\sum_{i=1}^n (\hat{Y}_i^c - \bar{Y}^c)}{\sum_{i=1}^n (Y_i^c - \bar{Y}^c)}$$

$$R_2^2 = \frac{\sum_{i=1}^n (\hat{Y}_i^r - \bar{Y}^r)}{\sum_{i=1}^n (Y_i^r - \bar{Y}^r)}$$

$$R^2 = (R_1^2 + R_2^2)/2$$

$$\hat{Y}_i^c = (\hat{Y}_L + \hat{Y}_U)/2$$

$$\bar{Y}^c = (\bar{Y}_L + \bar{Y}_U)/2$$

$$\hat{Y}_i^r = (\hat{Y}_U - \hat{Y}_L)/2$$

$$\bar{Y}^r = (\bar{Y}_U - \bar{Y}_L)/2$$

2.2.7 Root Mean Square Error (RMSE)

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

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

e_t - residual at t^{th} observation and n - overall count of observations.

3. RESULTS AND DISCUSSION

Data analysis was done mainly using R software package [16]. The distributions of lower limits and upper limits of the dengue counts in Colombo is given in Figure 1. More variability in upper bounds of monthly reported dengue cases can be seen than that of lower bounds. Reported lower and upper bounds in latter months were higher than that of former months.

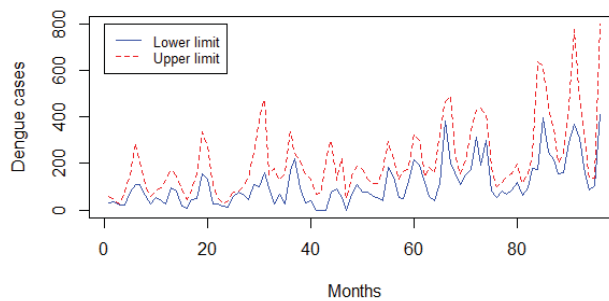


Figure 1. Distribution of lower limits and upper limits of dengue counts

Lower limits and upper limits of monthly maximum temperature indicated some cyclic variations which represents in the following figure (Figure 2).

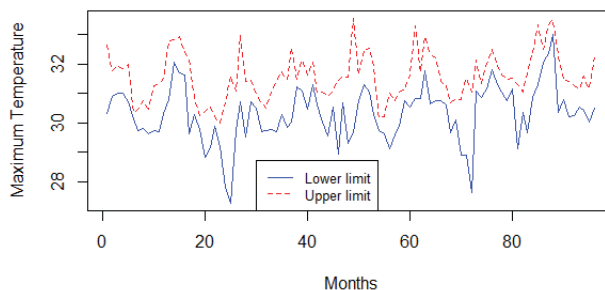


Figure 2. Distribution of lower limits and upper limits of maximum temperature

Figure 3 displays seasonal variations in lower limits as well as in upper limits of monthly minimum temperature in Colombo. Minimum and maximum of the lower bound are 22 and 26.7 whereas that for upper bound are 23.5 and 28.

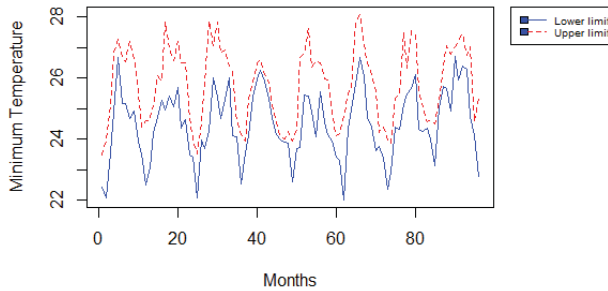


Figure 3. Distribution of lower limits with upper limits of minimum temperature

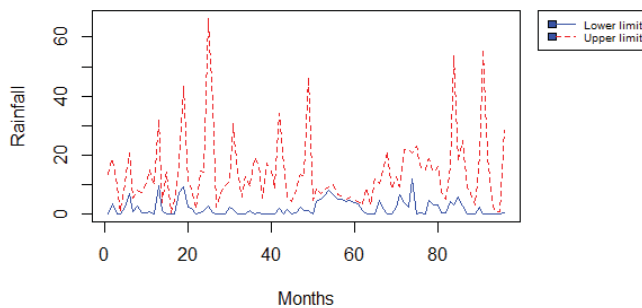


Figure 4. Distribution of lower limits with upper limits of rainfall

The distribution of lower limits of monthly rainfall is always fluctuated around zero but variability is high in upper bounds of monthly rainfall creating wider ranges within lower and upper intervals which can be seen in Figure 4.

Results of cross-correlation analysis are summarized in Table 1. Cross-correlation analysis revealed that 10 weeks leading rainfall and immediate minimum and maximum temperature effect on reported dengue cases. Therefore, 10-week lag period for rainfall with no lags for temperature variables modelled with reported dengue cases in each of the three methods; CM, CRM and CCRM. Data selection for the model development was from January 2009 to December 2015. Data from January to December in 2016 were used for validation of three models. The estimated coefficients of regression lines in each of the three models is shown in Table 2 and summary measures in Table 3.

Table 1: Cross-correlations of variables

Lag in weeks	Cross-correlations		
	Rainfall and Dengue	Maximum Temperature and Dengue	Minimum Temperature and Dengue
12	0.188	0.258	0.120
11	0.217	0.224	0.124
10	0.279	0.172	0.120
9	0.256	0.131	0.131
8	0.219	0.097	0.148
7	0.169	0.057	0.125
6	0.102	0.035	0.127
5	0.030	0.031	0.123
4	0.034	0.042	0.136
3	0.075	0.035	0.129
2	0.126	0.029	0.124
1	0.141	0.080	0.115
0	0.161	0.301	0.146

Table 2: The estimated coefficients of regression lines in CM, CRM and CCRM

	CM	CRM		CCRM	
		Center	Range	Center	Range
Intercept	42.39	42.39	107.67	42.39	77.65
$\hat{\beta}_0$	-2.75	-2.75	-20.14	-2.75	0.00
$\hat{\beta}_1$	5.76	5.76	-6.85	5.76	0.00
$\hat{\beta}_2$	4.79	4.79	2.08	4.79	1.66

Table 3: Summary measures in CM, CRM and CCRM

	RMSE	Lower	Upper	R-Squared	Standard Deviation	
					Center	Range
CM	71.29		118.31	0.40	89.09	-
CRM	69.02		113.73	0.41	89.09	73.62
CCRM	69.33		114.09	0.44	89.09	75.41

Highest RMSE (both lower and upper) reported in the CM. Although CRM reported lowest RMSE (both lower and upper), summary results of CRM and CCRM indicate that there is no significance difference. Predicted lower and upper limits from each of the models with actual bounds are shown in Figure 5-7.

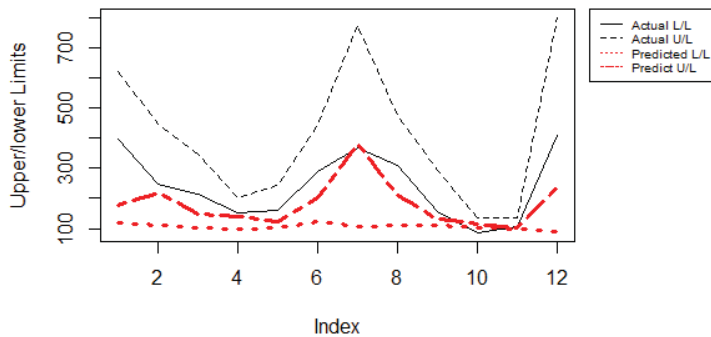


Figure 5. Actual with Forecasted Lower/Upper boundaries from CM

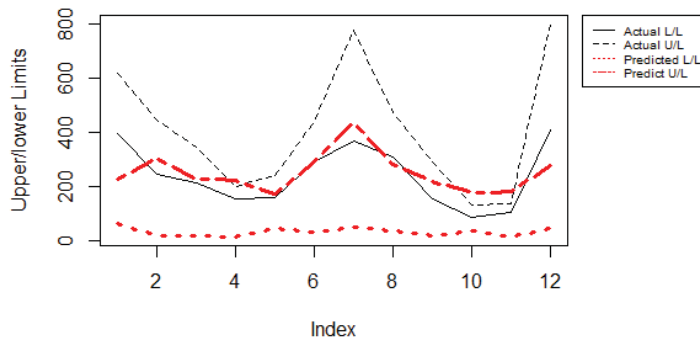


Figure 6. Actual with Forecasted Lower/Upper boundaries from CRM

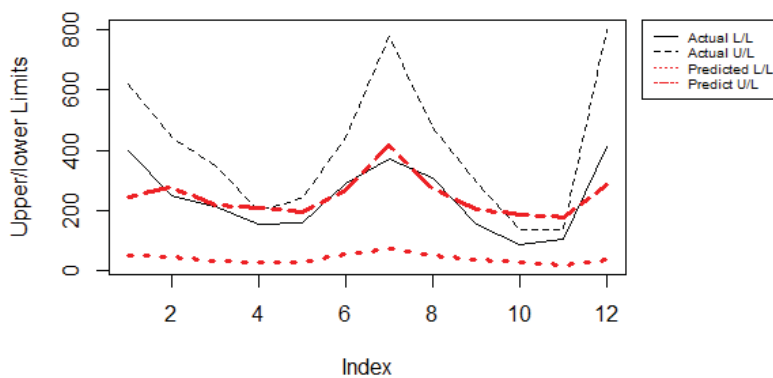


Figure 7. Actual with Forecasted Lower/Upper boundaries from CCRM

The predicted upper boundaries from CM is lower than the actual lower boundaries of dengue counts. Further, highest RMSE reported in CM. Hence, it is not suitable for predictions. Validation results from CRM and CCRM were close to each other but more smoothness can be seen in CCRM validation plot.

In this study, dengue counts were modeled using only three climatic factors such as maximum temperature, minimum temperature and rainfall. Not only climatic variables but also other variables such as urbanization, mobility patterns and random human effects can influence the dengue distribution. All of the significant external factors that can be associated with dengue distribution should be identified through a comprehensive study and those factors should be included into the model with the purpose of improving the precision/accuracy of prediction of the dengue, which is in interval scale.

4. CONCLUSION

This study successfully models the monthly interval valued dengue counts together with interval valued maximum (extreme) temperature, minimum (lowest) temperature and rainfall (amount of rain water) in Colombo, Sri Lanka with the aim of predicting the dengue cases in intervals. Predictions in intervals are important because exact predictions may not necessary in taking actions in controlling the dengue. Particularly, three interval valued regression procedures, namely, CM, CRM and CCRM were fitted on available interval valued data. Results revealed that among the CM, CRM and CCRM, either CRM or CCRM can be used to predict monthly interval-valued dengue cases in Colombo, Sri Lanka. The predicted values may be useful in taking controlling actions to manage the reported dengue cases in Colombo, Sri Lanka. Further, the CM, CRM and CCRM are based on ordinary least squares method. Authors are willing to implement other least squares techniques to handled interval valued data in modelling multiple regressions as a future study.

5. ACKNOWLEDGMENT

The grant number RPHS/2016/D05 of National Science Foundation (NSF), Sri Lanka was supported to conduct this research.

6. REFERENCES

1. World Health Organization, Dengue and Severe Dengue, retrieved from <http://www.who.int/mediacentre/factsheets/fs117/en/>.
2. Epidemiology Unit, Ministry of Healthcare and Nutrition, Sri Lanka, Dengue Update, retrieved from <http://www.epid.gov.lk>
3. Lai Y.H., 2018, The climatic factors affecting dengue fever outbreaks in southern Taiwan: an application of symbolic data analysis, *Biomed Eng Online*, **17(2)**, 148, doi:10.1186/s12938-018-0575-4
4. Cheong Y.L., Burkart K., Leitão P.J., Lakes T., 2013, Assessing weather effects on dengue disease in Malaysia, *Int J Environ Res Public Health*, **10(12)**, 6319–6334. doi: 10.3390/ijerph10126319.

5. Prakash T.G.S., Ranasinghe H.S., Karunadasa S., 2018, Dengue prevalence as an evidence of Climate change in Sri Lanka, *International Journal of Multidisciplinary Studies*, **5(2)**,23-30.
6. Wickramaarachchi W.P.T.M., Perera S.S.N., Jayasinghe S., 2015, Modelling and Analysis of Dengue Disease Transmission in Urban Colombo: A Wavelets and Cross Wavelets Approach, *Journal of National Science Foundation, Sri Lanka*, 43(4), 337-345.
7. Hii Y.L., Zhu H., Ng N., Ng L.C., Rocklöv J., 2012, Forecast of dengue incidence using temperature and rainfall, *PLOS Negl Trop Dis.*, doi: 10.1371/journal.pntd.0001908.
8. Promprou S., 2013, Multiple Linear Regression Model to Predict Dengue Haemorrhagic Fever (DHF) Patients in Kreang Sub-District, Cha-Uat District, Nakhon Si Thammarat, Thailand, *Journal of Applied Sciences Research*, 9(12), 6193 -6197.
9. Azman N.N.B., Karim S.A.B.A.,2018, Assessing Climate Factors on Dengue Spreading in State of Perak, *IOP Conf. Series: Journal of Physics*, **1123**, doi :10.1088/1742-6596/1123/1/012026.
10. Billard, L. & Diday, E., 2000, Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.), *Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies*, Namur, Springer, 369–374.
11. Lima, E.D. & de Carvalho, F.D.T., 2008, Centre and Range method for fitting a linear regression model to symbolic interval data, *Computational Statistics & Data Analysis*, **52(3)**, 1500–1515.
12. Lima, E.D. & de Carvalho, F.D.T., 2010, Constrained linear regression models for symbolic interval-valued variables, *Computational Statistics & Data Analysis*, **54(2)**, 333–347.
13. Draper, N.R. and Smith, H., 1998, *Applied Regression Analysis*. 3rd Edition, Wiley, New York. <https://doi.org/10.1002/9781118625590>.
14. Magda, M.M., 2018, Analysis of Multiple Linear Regression models using Symbolic Interval valued Variables. *International Journal of Applied Mathematics & Statistical Sciences*, **7(2)**, 33-44.
15. Xun Wang, Shoumei Li, Thierry Denoeux, 2015, Interval-Valued Linear Model. *International Journal of Computational Intelligence Systems*, Atlantis Press, **8(1)**, 114-127.
16. R Core Team, 2017, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
17. Peiris H.O.W., Chakraverty S., Perera S.S.N., Ranwala S.M.W., 2018, Novel Interval Multiple Linear Regression Model to assess the Risk of Invasive Alien Plant Species, *JSc EUSL*,**9(1)**,12-30.
18. Billard L, Diday E., 2007, *Symbolic data analysis: conceptual statistics and data mining*, New York: Wiley-Interscience.