Conference Paper No: PF-03

Fuzzy linear regression approach for modelling dengue disease transmission in Colombo

W. S. R. Fernando[1*] and A. M. C. H. Attanayake[1]

[1]Faculty of Science, University of Kelaniya, Sri Lanka
fernandosara1997@gmail.com*

Abstract

Dengue fever is caused by the dengue virus and is primarily transmitted through the bites of Aedes mosquitoes. Dengue fever is a significant public health concern in Colombo, Sri Lanka, with recurring outbreaks affecting a large number of individuals. Understanding the factors influencing dengue transmission is crucial for effective disease control and prevention. This study aims to explore a fuzzy regression approach for modelling dengue disease transmission in Colombo. As fuzzy linear regression incorporates uncertainty and imprecision into the modelling process, it has an advantage over conventional regression techniques. A fuzzy linear regression model was developed using various climate predictors, such as rainfall, relative humidity, wind speed, and temperature. Among these Climate variables, relative humidity and rainfall, are found to play a crucial role in mosquito breeding and the subsequent spread of the dengue virus. The fuzzy linear regression model is used to assess the relationships between these predictors and dengue transmission rates in Colombo. The suggested fuzzy linear regression model's evaluation criteria were done using R^2, Root Mean Squared Error, and Mean Absolute Error values. This study provides insights into the relationship between climatic factors and dengue transmission in Colombo by utilizing the above regression model. Climate variables, such as relative humidity and rainfall, are found to play a crucial role in mosquito breeding and the subsequent spread of the dengue virus. The findings highlight the importance of considering climatic predictors when developing dengue prevention strategies, particularly in urban environments like Colombo.

Keywords

Dengue fever, Fuzzy linear regression, Mean Absolute Error, R^2, Root Mean Squared Error.

Introduction

In recent years, dengue, a viral illness caused by mosquitoes, has expanded quickly throughout the entire world. It has been suggested that the dengue virus originated in Africa and that the slave trade played a role in its global spread (Sirisena & Noordeen, 2014). According to Withanage et al. (2018), dengue is primarily found in tropical and subtropical parts of the world, home to about 3.9 billion people. This is spread mostly by the Aedes aegypti and Aedes albopictus vectors. They are both highly domesticated urban mosquitoes that like to coexist with people in and around their homes, feed on them, and lay their eggs in small containers. In most tropical nations worldwide, dengue significantly burdens human populations, health, and economic systems compared to other diseases and their impacts. Over the past 50 years, dengue cases have dramatically increased. According to the American Centers for Disease Control and Prevention (CDC), up to 400 million people throughout the world contract an infection every year

(Amerasinghe et al., 1995). Over the past six years, dengue cases in the Colombo district have made up around 25% of all cases nationwide. Using climate data from the Colombo district, Leslie Chandrakantha (Chandrakantha, 2019) built a risk prediction model for dengue transmission. When building the model, a logistic regression technique was applied. Overall, the findings indicated that the only significant factor influencing the chance of experiencing increased dengue occurrences was rainfall. Using Bayesian Poisson spatial regression Thipruethai Phanitchat, et al. (Phanitchat et al., 2019) analysed Spatial and temporal patterns of dengue incidence in northeastern Thailand from 2006–2016. Over the course of this study, there was an increase in the number of dengue cases recorded in older age groups. The maximum ambient temperature was positively correlated with dengue incidence and was extremely seasonal. Climate-related factors, however, did not fully account for the regional variation of dengue in the province. Numerous types of research, like the ones described above, have demonstrated a connection between climatic variables and dengue incidences. Therefore, a risk prediction model based on climatic variables would help lower illnesses by eliminating mosquitoes that carry the virus. The current study is novel because there are limited applications of fuzzy linear regression in dengue transmission. In Sri Lanka, the spread of dengue is a significant public health issue. Dengue cases have been steadily rising in Sri Lanka over time. Moreover, no commercially available vaccination may be used to prevent dengue. Preventive interventions can work more effectively if there is a reliable early warning system. Determining the environmental and climatic elements that may be responsible for the rise in dengue cases and the resulting threat to human health is therefore vital. This study aims to model and forecast dengue disease in Colombo using fuzzy linear regression approach to forecast dengue cases. Generally, the majority of dengue cases are reported in the Colombo district and hence, the study focuses on Colombo. As fuzzy linear regression incorporates uncertainty and imprecision into the modelling process, it has an advantage over conventional regression techniques. Further, it can be given directions to the relevant authority for more attention to control the areas which have more dengue patients.

Methodology/materials and methods

Data Description

The monthly dengue disease incidence data is collected from the Epidemiology Unit, Ministry of Health, Sri Lanka and monthly meteorological data were obtained from the Department of the Census and Statistics in the Colombo district from 2010 to 2022. The data is further split into training and testing data sets, with training data being used to train the regression model.

Fuzzy Linear Regression

Fuzzy linear regression is a variant of linear regression that incorporates fuzzy logic concepts to handle uncertain or imprecise data. It extends traditional linear regression by allowing for fuzzy membership functions to represent the uncertainty associated with the input and output variables. Let X be a universal set, a fuzzy set $\widetilde{\mathrm{A}}$ of X is defined by its membership function $\mu_{\widetilde{\mathrm{A}}}: \mathrm{X} \rightarrow [0,1]$. The universal set X is generally assumed to be the set of real numbers (Buckley & Jowers, 2008).The fuzzy number $\widetilde{\mathrm{A}}$ is called LR –fuzzy number if it has the membership function as follows:

$$\mu_{\widetilde{A}}(x) = \begin{cases} L(x), & \text{if } x \in [a,b) \\ 1, & \text{if } x \in [b,c] \\ R(x), & \text{if } x \in (c,d] \\ 0, & \text{otherwise} \end{cases} \quad (01)$$

for all $x \in \mathbb{R}$ and $a < b < c < d$ where L, R: $[0,1] \to [0,1]$ are two shape functions such that $R[0] = L[0] = 0$ and $R[1] = L[1] = 1$. The support and the core of LR-fuzzy number $\widetilde{A}$ are closed intervals, i.e., $[a, b]$ and $[b, c]$, respectively. If $L[x] = \frac{x-b}{b-a}$, $R[x] = \frac{d-x}{d-c}$ and $b = c$, then LR fuzzy number $\widetilde{A}$ is called triangular fuzzy number (TFN). In general, TFN $\widetilde{A}$ can be denoted by $\widetilde{A} = (a, b, d)_T$ with the membership function as:

$$\mu_{\widetilde{A}}(x) = \begin{cases} \frac{x-b}{b-a}, & \text{if } x \in [a,b) \\ \frac{d-x}{d-c}, & \text{if } x \in [b,d] \\ 0, & \text{otherwise} \end{cases} \quad (02)$$

The goal of fuzzy linear regression is to find the best linear relationship between the fuzzy input variables and the fuzzy output variable. This involves estimating the parameters of the linear regression model, such as the intercept and coefficients, while considering the uncertainty associated with the fuzzy membership functions (Buckley & Jowers, 2008). The fuzzy linear regression model typically follows a similar form to traditional linear regression, but with fuzzy membership functions involved. The model can be represented as:

$$\hat{y}_i = \beta_0 + \beta_1(x_{i1}) + \cdots + \beta_p(x_{ip}) \quad (03)$$

In this equation, $\hat{y}_i$ is the predicted value of the fuzzy output variable for the i −th data point. x_{i1}, x_{i2}, ..., x_{ip} represent the fuzzy input variables associated with the i −th data point, each with their respective fuzzy membership functions. $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients or weights of the linear regression model. To formulate the above FLRM in Equation (03) can be presented according to the following steps (Rattanalertnusorn et al., n.d.).

Step 1- Defined LR-fuzzy numbers in the fuzzy model. Let $\tilde{y}_i$ and $\tilde{x}_{ij}$ are TFNs, which are denoted by

$$\tilde{y}_i = (y_{il}, y_{im}, y_{iu})_T \text{ and } \tilde{x}_{ij} = (x_{i1l}, x_{i1m}, x_{i1u})_T;\ j = 1, .., k \quad (04)$$

Also the membership function defined as:

$$\mu_{\tilde{y}_i}(y) = \begin{cases} \frac{y-y_{im}}{y_{im}-y_{il}}, & y_{il} < y < y_{im} \\ \frac{y_{iu}-y}{y_{iu}-y_{im}}, & y_{im} < y < y_{iu} \\ 0, & \text{otherwise} \end{cases} \quad (05)$$

And

$$\mu_{\tilde{x}ij}(x) = \begin{cases} \frac{x-x_{ijm}}{x_{ijm}-x_{ijl}}, & x_{ijl} < x < x_{ijm} \\ \frac{x_{iju}-x}{x_{iju}-x_{ijm}}, & x_{ijm} < x < x_{iju} \\ 0, & \text{otherwise} \end{cases} \quad (06)$$

Step 2- Transform the fuzzy components into crisp number by using the centroid formulae (Grzegorzewski, 2010) and obtain the conventional linear regression model.

Step 3- Formulated fuzzy linear regression model in Eq. (03) by using fuzzy output data and fuzzy input data are TFNs. Thus, the Eq. (03) can be rewritten by,

$$(y_{il}, y_{im}, y_{iu})_T = \beta_0 + \beta_1(x_{i1l}, x_{i1m}, x_{i1u})_T + \cdots + \beta_k(x_{ikl}, x_{ikm}, x_{iku})_T \quad (07)$$

Step 4- Fuzzy estimators were obtained using fuzzy least squares estimator vector,

$$\hat{\beta}_F = [x_l^t x_l + x_m^t x_m + x_u^t x_u]^{-1}[x_l^t x_l + x_m^t x_m + x_u^t x_u] \quad (08)$$

Results and Discussion

The descriptive statistics of the research variable includes the response variable and the explanatory variables, are given in Table 1.

Table 1. *Descriptive analysis of response variable and explanatory variables*

Variable	Minimum	Maximum	Mean	Standard deviation
Dengue Patients	30	3620	1013.1	744.1
Rainfall	0.1	971.5	211.6	168.7
Relative Humidity	71	86	80.694	3.234
Wind Speed	1.5	8	4.497	1.314
Temperature	26.55	29.9	0.0563	0.683

The study of the exploratory data showed that each climate factor affected dengue incidences differently. The impact of climate was anticipated to become apparent one or two months later because it takes time for an egg to mature into an adult mosquito (Nakhapakorn & Tripathi, 2005). Due to this, the climatic data with lags of two, five, and three months were used to model the Dengue Patient variable. For two lag-month data, the Pearson correlation coefficient between dengue incidences and rainfall was 0.357 ($p < 0.05$), indicating a significant positive association. For five lag-months data, the correlation coefficient between dengue incidences and relative Humidity was -0.409 ($p < 0.05$), indicating a significant negative association. Similarly for three lag-month data, the correlation coefficient between dengue incidences and wind speed was -0.237 ($p < 0.05$), indicating a significant negative association, and for two lag-month data, correlation coefficient between dengue incidences and temperature was 0.313 ($p < 0.05$), indicating a significant positive association.

Model: -Fuzzy Linear Regression

Table 2. *Parameters of classical regression model using crisp data*

Coefficient	Estimate	P-Value
Intercept (β_0)	4144.8238	0.003486
RainFall_Lag2 (β_1)	1.1536	0.000831
Relative Humidity_lag5 (β_2)	-41.9266	0.014495

Only the RainFall_Lag2 and Relative Humidity_lag5 parameters were significant under the 5% level of significance. The FLRM can be formulated as,
$(y_{il}, y_{im}, y_{iu})_T = \beta_0 + \beta_1(x_{i1l}, x_{i1m}, x_{i1u})_T + \beta_2(x_{i2l}, x_{i2m}, x_{i2u})_T$ Fuzzy estimators were obtained using fuzzy least squares estimator vector,

$$\hat{\beta}_F = [x_l^t x_l + x_m^t x_m + x_u^t x_u]^{-1}[x_l^t x_l + x_m^t x_m + x_u^t x_u]$$

Thus, $\hat{\beta}_{F0} =- 337.206$ $\hat{\beta}_{F1} = 1.47$ $\hat{\beta}_{F2} = 12.61$

Since $\hat{\tilde{y}}_i$ is a triangular fuzzy number and its membership function is,

$$\mu_{\hat{\tilde{y}}_i} = \begin{cases} \frac{\hat{y}-\hat{y}_{im}}{\hat{y}_{im}-\hat{y}_{il}}, & \hat{y}_{il} < \hat{y} < \hat{y}_{im} \\ \frac{\hat{y}_{iu}-\hat{y}}{\hat{y}_{iu}-\hat{y}_{im}}, & \hat{y}_{im} < \hat{y} < \hat{y}_{iu} \\ 0, & \text{otherwise} \end{cases}$$

where, $\hat{y}_{il} = -337.206 + 1.47x_{i1l} + 12.61x_{i2l}$

$\hat{y}_{im} = -337.206 + 1.47x_{i1m} + 12.61x_{i2m}$

$\hat{y}_{iu} = -337.206 + 1.47x_{i1u} + 12.61x_{i2u}$

Thus, the fuzzy output obtained as,

$$\hat{\tilde{y}}_i = -33.206 + 1.47(x_{i1l}, x_{i1m}, x_{i1u})_T + 12.61(x_{i2l}, x_{i2m}, x_{i2u})_T$$

This fuzzy model suggests that the predicted value $\hat{\tilde{y}}_i$ is a linear combination of the input variables x_{i1} and x_{i2}, where each input variable is assigned a membership value (l, m, u) indicating its degree of membership in a fuzzy set. The model uses the weights 1.47 and 12.61 to scale the contribution of each fuzzy input variable and adds them together with the intercept term (-33.206) to estimate the output $\hat{\tilde{y}}_i$.

Model Performance

Table 3. *Model performance of fuzzy linear regression*

Measure	Upper value	Middle Value	Lower value
R^2	31.7%	35.6%	39.2%
RMSE	1074	1092	936
MAE	771	785	640

Model performance of the Fuzzy linear Regression displays in Table 3, shows that R^2 values range from 31.7% to 39.2%, RMSE range from 936 to 1092, and MAE range from 640 to 785.

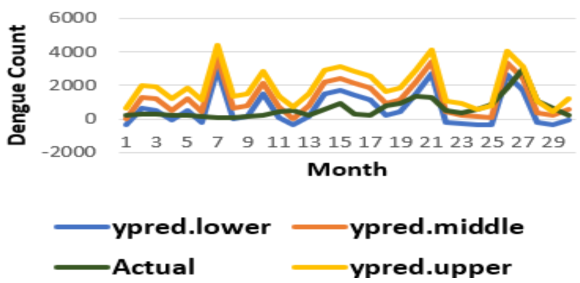

Figure 1. *Actual vs Fitted Values of Fuzzy linear Regression Model*

Actual values of the dengue count are between the upper bound and lower bound, as seen by the actual vs. fitted values for the test set. Fuzzy linear regression has the benefit of allowing lower and upper bounds to be set for the outcomes that are predicted. Fuzzy linear regression can deliver a range of potential outcomes rather than a single point estimate by adding fuzzy sets and membership functions. This lower and upper bound benefit enables a more thorough knowledge of the uncertainty related to the forecasting, improving risk management and decision-making processes. The only significant factor that increased dengue occurrences, according to previous studies, was rainfall. However, using a fuzzy linear regression model, it was discovered that both rainfall and relative humidity increased dengue cases. The study solely focuses on climate factors and does not take into account non-climatic factors, such as demographics, population immunity, socio-economic structure, availability of affordable public health facilities, and other environmental modification initiatives. Another limitation of this study is that a fixed validation size was used during the model evaluation process. The study is limited only to the Colombo district to analyse monthly climatic data and dengue incidence data. Future studies could expand the scope of analysis to cover more geographically diverse regions of Sri Lanka and use daily or weekly climate data for better accuracy. Additionally, vulnerable groups such as patients' age, gender, health status, and occupation could be taken into account to strengthen the surveillance system and facilitate more effective planning and preparation to prevent future outbreaks. In future studies, it would be valuable to explore alternative techniques for model evaluation, such as cross-validation and random sampling.

Conclusion

Dengue is a significant public health issue not only in Colombo but also in other regions of Sri Lanka. The objective of this study is to model and forecast dengue disease in Colombo using fuzzy regression approach and forecasting dengue cases based on climatic parameters. The study used monthly data from 2010 to 2022. The study found a significant correlation between meteorological parameters (rainfall and relative humidity) and dengue incidences. The study of the Fuzzy Linear Regression model's performance shows R^2 values spanning 31.7% to 39.2%, Root Mean Squared Error values between 936 and 1092, and Mean Absolute Error values range from 640 to 785. The R^2 value's potential range can be increased by including more explanatory factors. The departments of public health, medical researchers, and health geography analysts can make use of these study's outcomes to implement the required precautions based on such forecasts. By accurately predicting the number of dengue patients based on meteorological variables, these stakeholders can take proactive steps to reduce the spread of dengue, such as implementing targeted vector control measures, increasing public awareness campaigns, and improving healthcare infrastructure in areas at high risk of dengue outbreaks. Overall, the insights provided by this research can lead to improved public health outcomes and a reduction in the burden of dengue fever in the Colombo district.

References

Amerasinghe, F. P., Theobald, A., Meigen, A., Meigen, A., Theobald, A., Dyar, C., Linnaeus, C., Theobald, F., Ludlow, H., Theobald, H., Leicester, M., Blanchard, M., Theobald, M., Theobald, O., Leicester, T., Theobald, T., Giles, T., Lynch, U., & Genus Topomyia, A. (1995). *ILLUSTRATED KEYS TO THE GENERA OF MOSQUITOES (DIPTERA: CULICIDAE) IN SRI LANKA* (Vol. 23, Issue 4).

Buckley, J. J., & Jowers, L. J. (2008). Fuzzy sets. In *Studies in Fuzziness and Soft Computing* (Vol. 222, pp. 9–28). https://doi.org/10.1007/978-3-540-76290-4_2

Chandrakantha, L. (2019). *Risk Prediction Model for Dengue Transmission Based on Climate Data : Logistic Regression Approach*. 272–283.

Grzegorzewski, P. (2010). Algorithms for trapezoidal approximations of fuzzy numbers preserving the expected interval. *Studies in Fuzziness and Soft Computing*, *249*, 85–98. https://doi.org/10.1007/978-3-642-10728-3_5

Nakhapakorn, K., & Tripathi, N. K. (2005). An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *International Journal of Health Geographics*, *4*, 13. https://doi.org/10.1186/1476-072X-4-13

Phanitchat, T., Zhao, B., Haque, U., Pientong, C., Ekalaksananan, T., Aromseree, S., Thaewnongiew, K., Fustec, B., Bangs, M. J., Alexander, N., & Overgaard, H. J. (2019). *Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006 – 2016*. *3*, 1–12.

Rattanalertnusorn, A., Phannarook, S., & Jaroengeratikun, U. (n.d.). *A method for formulating fuzzy linear regression model and estimating the model parameters Parameter estimation in fuzzy linear regression model View project A method for formulating fuzzy linear regression model and estimating the model parameters*. *2015*, 49. https://www.researchgate.net/publication/275275247

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: IJID: Official Publication of the International Society for Infectious Diseases*, *19*, 6–12. https://doi.org/10.1016/j.ijid.2013.10.012

Withanage, G. P., Viswakula, S. D., Nilmini Silva Gunawardena, Y. I., & Hapugoda, M. D. (2018). A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. *Parasites & Vectors*, *11*(1), 262. https://doi.org/10.1186/s13071-018-2828-2.