Statistical Modelling of Dengue Cases in India during the period 2020-2022 : A Data Science Perspective

NAVAMI S (22MBS0003)
Post Graduate Student
School of Advanced Sciences
Vellore Institute of Technology, Vellore, India
Orcid ID: 0009000876206829
navami.s2022@vitstudent.ac.in

VRINDA S NAIR (22MBS0024)
Post Graduate Student
School of Advanced Sciences
Vellore Institute of Technology, Vellore, India
Orcid ID :0009000190314358
vrinda.snair2022@vitstudent.ac.in

NEDUNURI SAI SRIVIDYA (22MBS0032)
Post Graduate Student
School of Advanced Sciences
Vellore Institute of Technology, Vellore, India
Orcid ID:0009000313288541
sai.srividya2022@vitstudent.ac.in

Under the guidance of
DR. JITENDRA KUMAR
Associate Professor
School of Advanced Sciences
Vellore Institute of Technology
Vellore,India
Orcid ID: 0000000219429670
jitendra.kumar@vit.ac.in

***Abstract*— Dengue is one of the fastest-growing infectious diseases in the world, and it is a major public health threat in many countries, including India. It is a viral infection that is transmitted to humans by an Aedes mosquito species. Around the world, dengue is prevalent in areas which are prone to waterlogging. It is also linked to specific climatic and weather circumstances, which are more common in tropic-like regions or zones. Dengue is considered a dangerous disease because of the abrupt reduction in platelet count of infected patients, which frequently results in death.**

The present work aims to reduce the menace of dengue cases in India and other parts of the world. The goal of the present work is first to identify district-level hotspots. Secondly, to develop a suitable epidemiological model. This can be achieved through the implementation of a classification, cluster logistic model, and cluster analysis. Scan statistics is employed to determine the hotspots by finding the logistic function, the risk value, and the p-value to construct a suitable model. Machine learning techniques are employed to achieve efficacy in respect to various other epidemiological models.

Keywords*— *Dengue, Hotspot, Negative binomial regression, Logistic regression, Scan Statistics.

I. INTRODUCTION

Worldwide, dengue poses a serious risk to public health, especially in tropical and subtropical areas. [1] Aedes mosquitoes are the main vector of transmission to humans. There are four different serotypes of the dengue virus, and those who contract them can experience a variety of symptoms, like minor flu-like symptoms, while others may be severe, which can be fatal and is characterized by plasma leakage, severe bleeding, and organ dysfunction. Through a variety of strategies, such as the use of repellents and mosquito nets, prevention initiatives aim to reduce exposure to mosquito bites and control mosquito populations. In addition to continuing research into vaccines and treatments being a priority in the fight against dengue and minimizing its impact on global health, public health initiatives and community involvement are crucial in the fight against this illness. [2] [1] [3] [4]

II. METHODOLOGY

The objective of this project is :

1. To identify the district hotspots of dengue with the sample data.

2. Predicting the number of dengue cases and deaths using a negative binomial regression model.

3. To develop a logistic model for the outbreak of dengue

The present work is based on comprehensive research schemes where various statistical methods and tools namely logistic regression (as a predictive model), Negative Binomial Regression, scan statistics for hotspot analysis and other descriptive measures for different level of characterization of different cases has been planned and executed to achieve the defined objectives.

For the purpose of computational efficacy, the present work is executed using MS Excel, MS Solver, R, SatScan, and Python.

A. Hotspot Analysis

A location that has a larger concentration of events than one would anticipate from a collection of occurrences scattered randomly is referred to as a hotspot. [2]The analysis of point distributions and spatial arrangements of points in space led to the development of hotspot identification. When comparing point density within a specific area and looking at point patterns, a complete spatial randomness model—also called a

homogeneous spatial Poisson process—explains a process in which point occurrences occur completely at random. The major three files needed for hotspot analysis are. They are as follows:

1. Case file, which contains information on the number of incidents in the area.

2. The population file contains information about the population in the area.

3. The Location file consists of the details of the longitude and latitude of the geographical area.

With the use of these three files, hotspot analysis will enable us to pinpoint the primary cluster, secondary cluster, tertiary cluster, etc. that is most likely to have the most events. And doing so assists in determining the severity and consistency of the clusters over time, which helps create plans of attack for the epidemics in those identified clusters. [3]

B. Negative Binomial Regression:

The dependant variable (Y) in negative binomial regression is an observed count with a negative binomial distribution, in contrast to regular multiple regression. Therefore, nonnegative numbers like 0, 1, 2, 3, and so forth are feasible possibilities for Y. In negative binomial regression, the dependent variable mean is calculated using the time t and k independent variables.

$$\mu_i = \exp\left(\ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right) \qquad (1)$$

The fundamental framework for an i-th observation in negative binomial regression model is,

$$\Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \qquad (2)$$

Negative Binomial regression belongs to the category of generalized linear models (GLMs) and is specifically applied when the dependent variable represents a count of events and the data exhibits overdispersion. Overdispersion occurs in count data when the mean is exceeded by the variance of the dependent variable.

C. Logistic Regression

A logistic regression is a statistical analytic technique that predicts a binary outcome, such as yes or no. It is a model that predicts a dependent variable in the data by analysing the correlation among one or more current independent variables. In machine learning, logistic regression has grown its significance as a tool. It makes it possible for algorithms to categorize incoming input according to historical data in machine learning applications. The algorithms improve in their ability to anticipate classes within data sets as more pertinent data becomes obatinable. Furthermore, logistic regression can assist in data preparation endeavors by arranging data sets into predetermined categories during the extract, transform, load (ETL) procedure. [4] [5] [6] [7]

$$y_i^{\wedge} = \prod(X) = \frac{1}{1 + e^{(-X\beta)}} \qquad (3)$$

III .Analysis & Results

A. Hotspot Analysis:

The utilization of a discrete scan statistic arises from the discrete nature of the data concerning dengue cases and fatalities. A Poisson-based model is utilized in which the distribution of occurrences within a specific location follows a Poisson distribution. The research concentrates on the states that are primarily impacted by geography. Therefore, a strict spatial analysis is carried out. A pure geographical analysis of dengue cases and deaths from 2020 to 2022 was conducted to identify clusters with high rates.

Table 1: Hotspot clusters in cases of dengue over the period (2020-2022) district-wise.

Year	Most likely cluster(cases)	Most likely cluster (deaths)
2020	Kannur	Kannur
2021	Firozabad	Bellary
2022	Fatehgarh Sahib	Bangalore Rural, Chikkaballapur, Krishnagiri, Mysuru

B. Negative Binomial regression:

A negative binomial regression model has been developed for both dengue cases and deaths respectively by taking the annual rainfall and area of the states as independent variables together for the years 2020,2021,2022. The dependent variable is the number of dengue cases and number of dengue deaths for the respective models. The following are the resulting models obtained with coefficients and various parameters like deviance, log likelihood etc. It is obtained that both the models are well developed since the ratio of the deviance with the degrees of freedom that is the difference between the total number of observations and number of parameters i.e.105-5=100 of both the models is greater than 1.

Negative Binomial Regression Model for Dengue Cases:
Generalized Linear Model Regression Results

Dep. Variable:	cases	No. Observations:	105
Model:	GLM	Df Residuals:	102
Model Family:	NegativeBinomial	Df Model:	2
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-960.37
Date:	Tue, 31 Oct 2023	Deviance:	307.19
Time:	19:30:43	Pearson chi2:	436.
No. Iterations:	10	Pseudo R-squ. (CS):	0.4085
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	8.8890	0.254	35.034	0.000	8.392	9.386
rainfall	-0.0006	0.000	-6.118	0.000	-0.001	-0.000
Area	3.759e-06	1.12e-06	3.350	0.001	1.56e-06	5.96e-06

Negative Binomial Regression Model for Dengue Deaths:
Generalized Linear Model Regression Results

Dep. Variable:	deaths	No. Observations:	105
Model:	GLM	Df Residuals:	102
Model Family:	NegativeBinomial	Df Model:	2
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-290.23
Date:	Tue, 31 Oct 2023	Deviance:	270.14
Time:	19:30:43	Pearson chi2:	379.
No. Iterations:	12	Pseudo R-squ. (CS):	0.3445
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.1054	0.288	7.311	0.000	1.541	2.670
rainfall	-0.0005	0.000	-3.617	0.000	-0.001	-0.000
Area	3.79e-06	1.19e-06	3.187	0.001	1.46e-06	6.12e-06

C. Logistic Regression

Accuracy: 0.90

	precision	recall	f1-score	support
0	0.80	1.00	0.89	8
1	1.00	0.85	0.92	13
accuracy			0.90	21
macro avg	0.90	0.92	0.90	21
weighted avg	0.92	0.90	0.91	21

Deviance: 3099.75
Number of Observations: 105

The threshold for the variable y outbreak has been taken on the basis of proportion of the number of dengue incidents with the annual rainfall. If the proportion is greater than 0.5 then it is classified as outbreak(1) otherwise 0.

The developed model has been giving an accuracy of 90% and the training and testing data has been taken in the ratio 80:20.

IV .CONCLUSION

According to hotspot analysis, the districts in Karnataka like Bellary, Bengaluru Rural, Chikkaballapur, Krishnagiri, and Mysuru have emerged as the primary regions for dengue-related deaths in the years 2021 and 2022. These findings indicate the inadequate treatment provided to patients suffering from dengue in these areas.

Using Negative binomial regression two predictive models for cases and deaths respectively have been developed which can be used to predict the cases and deaths and the logistic regression equation is used to classify the outbreak areas based on these predictions.

$$y_i^{\wedge} = \prod(X) = \frac{1}{1 + e^{-(1.3609 - 0.66x_1 + 2.8347x_2 + 0.779x_3 + 0.5681x_4)}}$$

Where x_1 represents annual rainfall, x_2 represents number of cases , x_3 represents number of deaths and x_4 represents the area of the state.

V .REFERENCES

[1] S. S. A. J. &. U. C. C. Nivedita Gupta, "Dengue in India," *Centenary review Article,* pp. 373-390, September 2012.

[2] B. R. K. J. K. Anjali, "Spatio-Temporal Aspect of Suicide and Suicidal Ideation: An Application of SaTScan to Detect Hotspots in Four Major Cities of Tamil Nadu," *Journal of Scientific Research of The Banaras Hindu University,* vol. 65, no. 9, 2021.

[3] "https://www.publichealth.columbia.edu/research/population-health-methods/hot-spot-spatial-analysis," [Online].

[4] R. M. I. M. A. D. I. L. M. S. P. B. D. S. U. M. M. N. K. Xing Yu Leung, "A systematic review of dengue outbreak prediction models: Current scenario and future directions.," *PLOS Neglected Tropical Diseases,* 2023.

[5] J. L. J. N.-T. N. V. ,. J. C. C. a. A. J. R.-M. Maritza Cabrera, *Dengue Prediction in Latin America Using Machine Learning and the One Health Perspective: A Literature Review,* Switzerland: Tropical Medicine and Infectious Disease, 2022.

[6] N. d. W. J. M. A. W. Simon Hales, "Potential effect of population and climate changes on global distribution of dengue fever : an empirical model.," *THE LANCET,* vol. 360, pp. 830-4, 2002.

[7] H. G. L. X. L. Y. a. J. D. Zhichao Li, "Improving Dengue Forecasts by Using Geospatial Big Data Analysis in Google Earth Engine and the Historical Dengue Information-Aided Long Short Term Memory Modeling," *Biology,* vol. 11, p. 169, 2022.

[8] O. M. E. S. E. A. B. E. N. M. Corey M. BenedumID, "Statistical modeling of the effect of rainfall flushing on dengue transmission in Singapore," *PLOS Neglected Tropical Diseases,* 2018.

[9] S. Polwiang, "The time series seasonal patterns of dengue fever and associated weather variables in Bangkok(2003-2017)," *BMC.*

[10] S. N. S. J. K. KhushbooSukhija, "Spatial Visualization Approach for Detecting Criminal Hotspots: An Analysis of Total Cognizable Crimes in the State of Haryana," in *2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT)*, 2017.