International
Journal
Of Advanced Research in Engineering& Management (IJAREM)
ISSN: 2456-2033 || PP. 01-08

New Covid Cases in India during Second Wave: Data Prediction using Probability Modelling

S. Sivamani[a]* and N. Abdul Nazar[a]
[a]*University of Technology and Applied Sciences, Salalah, Oman*

Abstract: In India, the first coronavirus disease (COVID) case was confirmed on 30/01/2020. During the first wave, the number of new cases increase and become minimum of 9102 on 25/1/2021. After fluctuations for 4 days, COVID cases start to increase for the second wave from 13044 on 30/01/2021. In this manuscript, mathematical models based on probability density function was proposed to predict new cases infected with COVID during the second wave in India and the models were validated using various error functions. Gaussian, fuzzy logic generalized membership, hyperbolic secant, Witch of Agnesi, logistic and random algebraic functions were used for probability modelling. Average of error (AE), sum of square of error (SSE), average of relative error (ARE), correlation coefficient (R), determination coefficient (R^2), adjusted determination coefficient (Adj. R^2), coefficient of variance (CV), Chi-square (χ^2), bias, Akaike information criterion (AIC), Bayesian information criterion (BIC)andAmemiya's prediction criterion **(**APC**)** were used for validating the models. From the results, not a single probability model could fit the new COVID cases data in India during the second wave.

Keyword: COVID, Modelling, Probability, Validation, Error functions

Introduction:

Coronavirus disease (COVID) has infected people in almost all parts of world (Ahmed et al., 2020). Hence, the World Health Organization (WHO) declared the outbreak a public health emergency of International concern on 30/01/2020, and a pandemic on 11/03/2020 (Bhattacharjee, 2020).India is not an exception for COVID infection. The first case of COVID was identified in India on 30/01/2020 (Andrews et al., 2020). The infection reached its peak on 16/09/2020 with new cases of 97894. Again, number of cases gradually declines to 9102 on 25/01/2021. COVID cases start to increase for the second wave from 13044 on 30/01/2021 after fluctuating for few days (Chatterjee et al., 2020). During the second wave, the number of new cases reached its highest at 414188 on 06/05/2021. In this study, mathematical modelling is proposed on new case prediction for second wave in India.

Modelling is anequation or a set of equations (models) that explains the behaviour of a system. In a simpler way, models are the mathematical representation of data models (Hovmand, 2003). Models are of two types – Theoretical and empirical models. Theoretical models explain the basic mechanism of the process whereas empirical models are used to forecasting (Leonidou and Katsikeas, 1996). The process of modelling requires both domain and mathematical knowledge. Domain knowledge is used for identification of problem statement, selection of significant predictor and response variables, and generation of experimental data relating independent and dependent variables.Mathematical knowledge is required for selection of appropriate models for the available or generated data (Peverly and Sumowski, 2012) .

For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since, for continuous distributions, the probability at a single point is zero, the function f(x) is often expressed in terms of an integral between two points, a and b (Dragulescu and Yakovenko, 2002).

$$\int_b^a f(x).dx = p(a \le x \le b)$$

where p is probability function. For a discrete distribution, the pdf is the probability that the variate takes the value x.

$$f(x) = p(x)$$

Panovska-Griffiths (2020) discussed the importance of modelling in understanding COVID spread, highlighted different modelling approachesand suggested that no one model can give all the answers.McBryde et al., (2020) demonstrated the pandemic potential of model projections based on theinfectiousness of virus, which guided the global response to and prepared countries for increases in hospitalisations and deaths.Schaback (2020) analysed the COVID-19 pandemic by comparably simple mathematical and numerical methods. The final goal was to predict the peak of the epidemic outbreak per country with a reliable technique. From the analysis of literature, probabilistic models have not yet been attempted to analyse the COVID outbreak. Hence, in the present study, mathematical models based on probability density function was proposed to predict new cases infected with COVID during the second wave in India and the models were validated using various error functions.

Methodology:

Work plan:

The work plan of the present research is shown in Figure 1.Data on new cases of COVID was collected from an authentic source. Based on the COVID trending, suitable mathematical functions are selected and fit the collected data to the model. As all the models do not show the goodness-of-fit for the data, the models showing better fitness will be selected based on error functions.

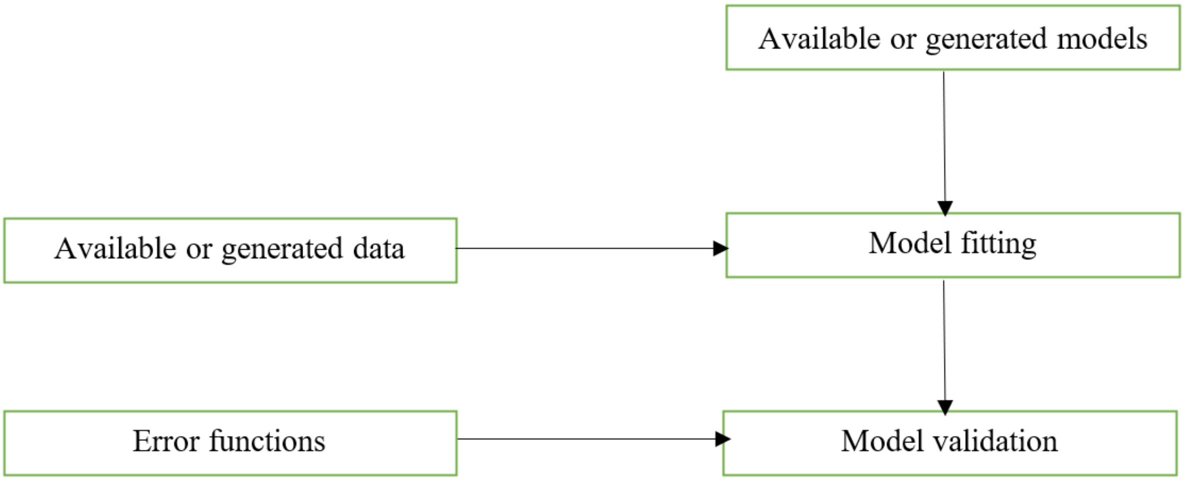

Figure 1. Work plan of the present investigation

Data collection:

In this investigation, Our World In Data (OWID) was used a source of information for COVID infected new cases in India from 30/01/2020 to 29/07/2021. The data was used without further modifications. The datasheet provides information on The International Organization for Standardization (ISO) code, continent, location, date, total cases, new cases, new cases smoothed, total deaths, new deaths, new deaths smoothed, total cases per million, new cases per million, new cases smoothed per million. Total deaths per million, new deaths per million, new deaths smoothed per million, reproduction rate, intensive care unit (ICU) patients, ICU patients per million, hospitalised patients, hospitalised patients per million, weekly ICU admissions, weekly ICU admissions per million, weekly hospitalised admissions, weekly hospitalised admissions per million, new tests, total tests, total tests per thousand, new tests per thousand, new tests smoothed, new tests smoothed per thousand, positive rate, tests per case, tests units, total vaccinations, people vaccinated, people fully vaccinated. New vaccinations, new vaccinations smoothed, total vaccinations per hundred, people vaccinated per hundred, people fully vaccinated per hundred, new vaccinations smoothed per million, stringency index Population, population density, median age, aged 65 older, aged 70 older, gross domestic product (GDP) per capita, extreme poverty, cardio vascular death rate, diabetes pre valence, female smokers, male smokers, hand washing facilities, hospital beds per thousand, life expectancy, human development index and excess mortality (Mathieu et al., 2021). The daily new cases of COVID infection data for India from 30/01/2020 is shown in Figure 2.

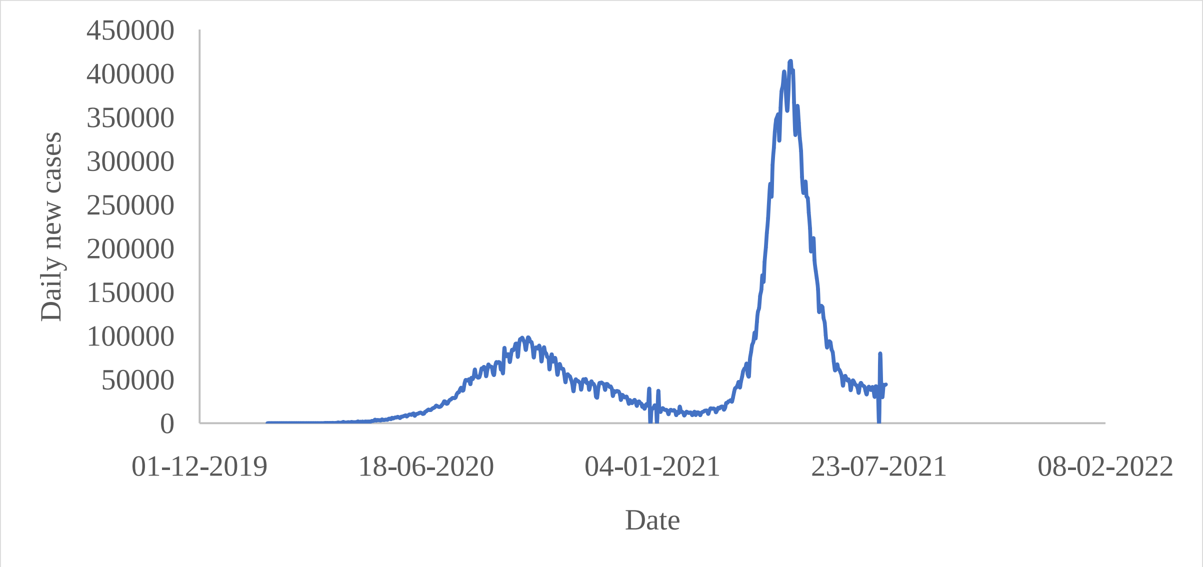

Figure 2. Daily new cases of COVID infection in India

Model selection:

Vast mathematical models are available to fit the data. An utmost care shall be taken in the selection of models.Model selection is an important criterion for model fitting process. Mathematical knowledge is required for the appropriate selection of models. From the Figure 2, it was inferred that the curve followed bell curve. Also, the probability density function follows bell curve. The functions that follow the bell curve are Gaussian, fuzzy logic generalized membership, hyperbolic secant, Witch of Agnesi, logistic and random algebraic functions (Kollu et al., 2012). The functions are listed in Table 1.

Table 1. Mathematical functions for probability density models

Probabilistic model	**Function**
Gaussian model	$\delta e^{((x-\mu)/\sigma)^2}$
Fuzzy logic generalized membership model	$\frac{1}{1+\left\|\frac{x-c}{a}\right\|^{2b}}$
Hyperbolic secant model	$\alpha\,\mathrm{sech}(\gamma x)$
Witch of Agnesi model	$\frac{d^3}{x^2+d^2}$
Logistic model	$\frac{\omega\, e^{\beta x}}{(1+e^{\beta x})^2}$
Random algebraic model	$\frac{p}{(1+x^2)^q}$

a, b, c, d, p, q, α,β, γ, δ, μ, ω and σ are model coefficients

Model fitting and validation:

The data collected from OWID for new COVID cases in India from 30/01/2021 was substituted in the probability density or bell-shaped models. As a result of model fitting, the values of mean, standard deviation, and constants were evaluated, and the predicted new cases were calculated from the same model. The deviation between actual and predicted new cases were calculated based on error functions. Average of error (AE), sum of square of error (SSE), average of relative error (ARE), correlation coefficient (R), determination coefficient (R^2), adjusted determination coefficient (Adj. R^2), coefficient of variance (CV), Chi-square (χ^2), bias, Akaike information criterion (AIC), Bayesian information criterion (BIC) andAmemiya's Prediction Criterion **(**APC**)**are used to evaluate the fitness of model (Sivamani et al., 2021). The functions are tabulated in Table 2.

Table 2. Equations for error functions

Error function	**Equation**
Average of error (AE)	$\frac{1}{n}\sum_{i=1}^{n}(Y_{ac} - Y_{pr})$
Sum of square of error (SSE)	$\sum_{i=1}^{n}(Y_{ac} - Y_{pr})^2$
Average of relative error (ARE)	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_{ac} - Y_{pr}\rvert}{Y_{ac}}$
Correlation coefficient (R)	$\sqrt{\frac{\sum_{i=1}^{n}\left(Y_{pr} - \bar{Y}_{pr}\right)^2}{\sum_{i=1}^{n}\left(Y_{pr} - \bar{Y}_{pr}\right)^2 + \sum_{i=1}^{n}(Y_{ac} - Y_{pr})^2}}$
Determination coefficient (R^2)	$\frac{\sum_{i=1}^{n}\left(Y_{pr} - \bar{Y}_{pr}\right)^2}{\sum_{i=1}^{n}\left(Y_{pr} - \bar{Y}_{pr}\right)^2 + \sum_{i=1}^{n}(Y_{ac} - Y_{pr})^2}$
Adjusted determination coefficient (Adj. R^2)	$1 - \left[(1 - R^2)\left(\frac{n-1}{n-p-1}\right)\right]$
Coefficientof variance (CV)	$\frac{\sigma}{\mu}x100$
Chi-square (χ^2)	$\sum_{i=1}^{n}(Y_{ac} - Y_{pr})^2/Y_{pr}$
Bias	$exp\left(\frac{1}{n}\sum_{i=1}^{n}ln\left(\frac{Y_{ac}}{Y_{pr}}\right)\right)$
Akaike information criterion (AIC)	$n\ln(SSE) - n\ln(n) + 2p$
Bayesian information criterion (BIC)	$n\ln(SSE) - n\ln(n) + p\ln(n)$
Amemiya's Prediction Criterion **(**APC**)**	$\frac{(n+p)}{n(n-p)}SSE$

Y_{ac} = actual number of new cases, Y_{pr} = Predicted number of new cases, $\bar{Y}_{pr}$ = Average of predicted number of new cases n = number of observations, p = number of constants in model.

Results and Discussion:

Figure 3 shows the actual new COVID cases in India during the second wave from 30/01/2021 to 29/07/2021 with reference to smoothed new cases. A fluctuation in number of new cases was observedfrom 30/01/2021 to the mid of March 2021. The increase was observed on 18/02/2021, 25/02/2021, 06/03/2021 at 13193, 14284 and 18754, respectively. The actual elevation started from 35871 on 17/03/2021. The increase in number of new COVID cases between 6th and 17th March is pathetic, i.e., from 18754 to 35871, almost doubled. Post 17/03/2021, the number of new cases doubled or more than doubled every 15 days. The number of new cases were 72330, 200739, 379308 and 414188 on 31/03/2021, 14/04/2021, 28/04/2021 and 06/05/2021, respectively.

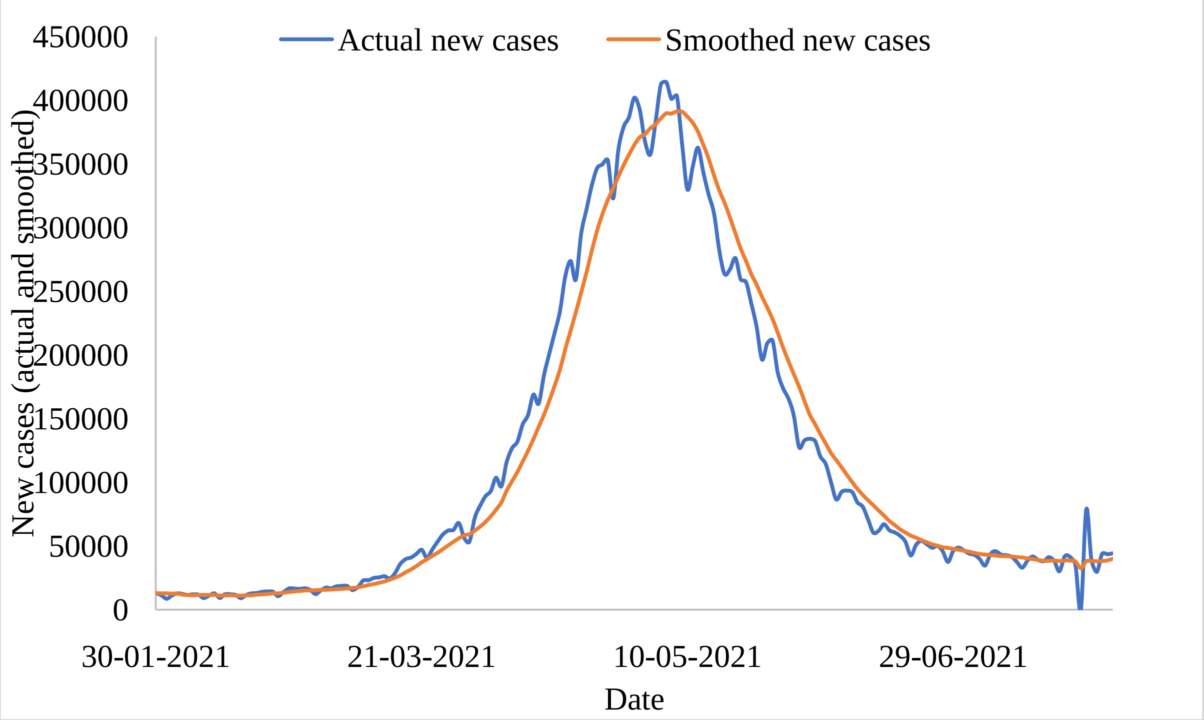

Figure 3. New COVID cases in India during the second wave from 30/01/2021 to 29/07/2021

After reaching the peak on 6th May 2021, the number of new cases start to decline to 362727, 211298 and 93463 on 12/05/2021, 26/05/2021 and 09/06/2021, respectively. Again, fluctuation was observed after 9th June 2021. It was 51667, 43393, 38164 and 44230, on 23/06/2021, 08/07/2021, 18/07/2021 and 29/07/2021, respectively.

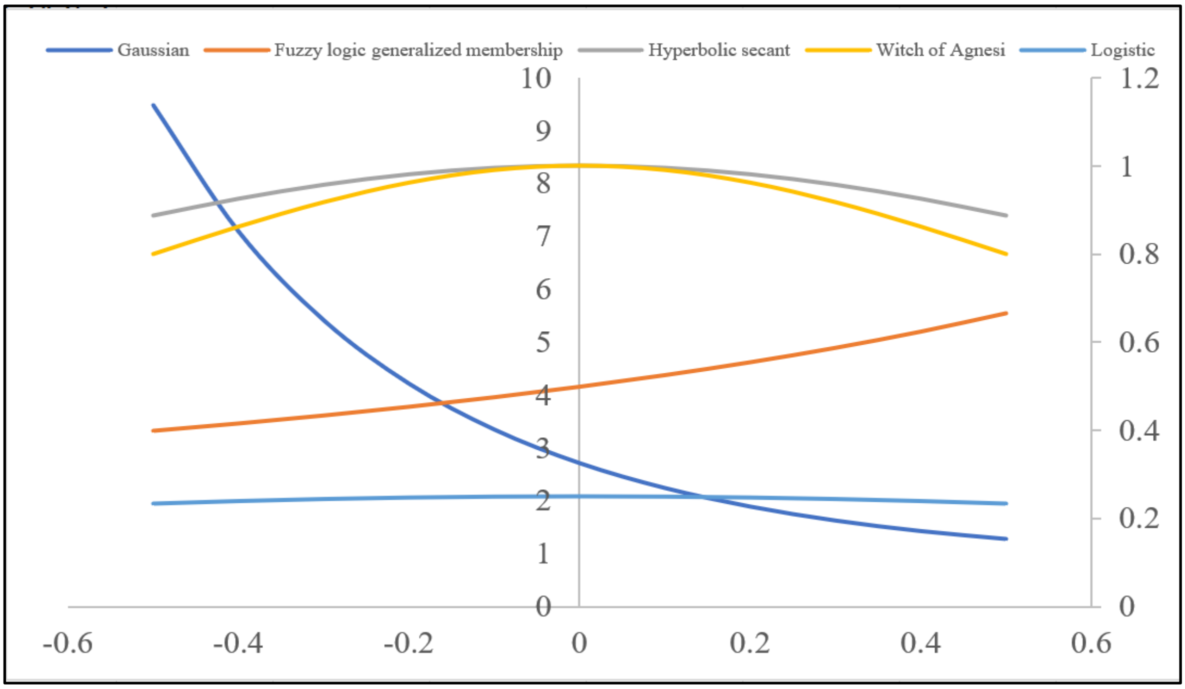

Figure 4. Graphical representation of various probabilistic models

The mathematical models employed for fitting the COVID data were Gaussian, fuzzy logic generalized membership, hyperbolic secant, Witch of Agnesi, logistic and random algebraic functions. Figure 4 represents the graphical representation of various probabilistic models. Gaussian function is the probability density function of the normal distribution. This is the archetypal bell-shaped function and frequently encountered in nature as a consequence of the central limit theorem. Fuzzy logic generalized membership function computes fuzzy membership values using a generalized bell-shaped membership function. Hyperbolic secant is the derivative of the Gudermannian function. Witch of Agnesi is the probability density function of the Cauchy distribution. This is also ascaled version of the derivative of the arctangent function. Logistic function is the derivative of the simple logistic model. This is a scaled version of the derivative of thehyperbolic tangent function.

Model fitting is a procedure that takes three steps: (i) A function or model is required that takes in a set of parameters and returns a predicted data set; (ii) Error functions are selected that provides a number representing the difference between actual and predicted data for any given set of model parameters. This is usually either the sum of squared error (SSE) or maximum likelihood; and (iii) The parameters that minimize this difference should be identified.

Average of error (AE), sum of square of error (SSE), average of relative error (ARE), correlation coefficient (R), determination coefficient (R^2), adjusted determination coefficient (Adj. R^2), coefficient of variance (CV), Chi-square (χ^2), bias, Akaike information criterion (AIC), Bayesian information criterion (BIC) andAmemiya’s Prediction Criterion **(**APC**)** are the error functions used to validate the models. Basically, error is the difference between actual and predicted values. The following criteria shall be followed for the best model: (i) Average of error, sum of square of error and average of relative error to be minimum or close to zero; (ii) Correlation, determination and adjusted determination coefficients to be close to one; (iii) Coefficient of variance to be less than 10%; (iv) Chi-square to be close to zero; (v) Bias to be close to one; (vi) minimum AIC, maximum BIC and minimum APC suit for the best model.

Table 3 shows the model coefficients and error functions for various probabilistic models. Model coefficients were evaluated using Goal Seek What-if-function and Solver add-in of Microsoft Excel 16.0. Goal Seek What-if-function is used to find coefficients in models involving single constant term. Solver add-in is used to evaluate multiple coefficients.

Table 3. Model coefficients and error functions for various probabilistic models

Error function	Models					
	Gaussian model	Fuzzy logic generalized membership model	Hyperbolic secant model	Witch of Agnesi model	Logistic model	Random algebraic model
Model coefficients	δ = 36.07, μ = 288888.6, and σ = 86101.87	a = 510226, b = 10000 and c = 10000	α = 4840026 and γ = 0.0001	d = 115610	β = 0.0001 and ω = 9943748.5	p = 117624 and q = 0.0001
Average of error (AE)	0	0	0	0	0	0
Sum of square of error (SSE)	2.67x10^{12}	2.67x10^{12}	2.67x10^{12}	2.67x10^{12}	2.67x10^{12}	2.67x10^{12}
Average of relative error (ARE)	2.55	2.55	2.55	2.55	2.55	2.55
Correlation coefficient (R)	-0.14	-0.14	-0.14	0.063	-0.14	-0.14
Determination coefficient (R^2)	0.02	0.02	0.02	0.004	0.02	0.02
Adjusted determination coefficient (Adj. R^2)	0.01	0.01	0.01	-0.001	0.01	0.01
Coefficient of variance (CV)	65.95	65.96	65.94	66.01	65.98	65.97
Chi-square (χ^2)	0	0	0	0	0	0
Bias	-0.60	-0.59	-0.59	-0.59	-0.60	-0.59
Akaike information criterion (AIC)	4232.01	4231.90	4234.05	4242.81	4234.04	4233.07
Bayesian information criterion (BIC)	154.71	154.24	159.82	163.88	160.49	159.63
Amemiya's Prediction Criterion **(APC)**	1.52x10^{10}	1.52x10^{10}	1.46x10^{10}	1.49x10^{10}	1.5x10^{10}	1.5x10^{10}

As per the given criteria, average of error approaches zero but sum of squared of error value is too large. Similarly, average of relative error is small enough, but correlation, determination and adjusted determination coefficients are far away from unity. Chi-square is zero but bias and coefficient of variance are not appropriate. Information criterion proposed by Akaike, Bayes' and Amemiya do not follow the criteria.To summarize, based on sum of squared of error, no single model is suitable for fitting the new COVID cases data in India during the second wave.

Conclusion:

The present research aimed to propose mathematical models based on probability density function to predict new cases infected with COVID during the second wave in India and to validate the models using various error functions. The mathematical models employed for fitting the COVID data were Gaussian, fuzzy logic generalized membership, hyperbolic secant, Witch of Agnesi, logistic and random algebraic functions. Average of error (AE), sum of square of error (SSE), average of relative error (ARE), correlation coefficient (R), determination coefficient (R^2), adjusted determination coefficient (Adj. R^2), coefficient of variance (CV), Chi-square (χ^2), bias, Akaike information criterion (AIC), Bayesian information criterion (BIC) andAmemiya's Prediction Criterion **(**APC**)** are the error functions used to validate the models. It could be concluded that not a single probability model could fit the new COVID cases data in India during the second wave.

References:

[1]. Ahmed, M. A., Jouhar, R., Ahmed, N., Adnan, S., Aftab, M., Zafar, M. S., & Khurshid, Z. (2020). Fear and practice modifications among dentists to combat novel coronavirus disease (COVID-19) outbreak. *International journal of environmental research and public health*, *17*(8), 2821

[2]. Andrews, M. A., Areekal, B., Rajesh, K. R., Krishnan, J., Suryakala, R., Krishnan, B., ... & Santhosh, P. V. (2020). First confirmed case of COVID-19 infection in India: A case report. *The Indian journal of medical research*, *151*(5), 490.

[3]. Bhattacharjee, S. (2020). Statistical investigation of relationship between spread of coronavirus disease (COVID-19) and environmental factors based on study of four mostly affected places of China and five mostly affected places of Italy. *arXiv preprint arXiv:2003.11277*.

[4]. Chatterjee, S., Sarkar, A., Chatterjee, S., Karmakar, M., & Paul, R. (2020). Studying the progress of COVID-19 outbreak in India using SIRD model. *Indian Journal of Physics*, 1-17.

[5]. Dragulescu, A. A., & Yakovenko, V. M. (2002). Probability distribution of returns in the Heston model with stochastic volatility. *Quantitative finance*, *2*(6), 443.

[6]. Hovmand, P. S. (2003). Analyzing dynamic systems: a comparison of structural equation modeling and system dynamics modeling. *Structural equation modeling: applications in ecological and evolutionary biology*, 212-234.

[7]. Kollu, R., Rayapudi, S. R., Narasimham, S. V. L., &Pakkurthi, K. M. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of energy and environmental engineering*, *3*(1), 1-10.

[8]. Leonidou, L. C., &Katsikeas, C. S. (1996). The export development process: an integrative review of empirical models. *Journal of international business studies*, *27*(3), 517-551.

[9]. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., ... &Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature human behaviour*, 1-7.

[10]. McBryde, E. S., Meehan, M. T., Adegboye, O. A., Adekunle, A. I., Caldwell, J. M., Pak, A., ... &Trauer, J. M. (2020). Role of modelling in COVID-19 policy development. *Paediatric respiratory reviews*, *35*, 57-60.

[11]. Panovska-Griffiths, J. (2020). Can mathematical modelling solve the current Covid-19 crisis?.

[12]. Peverly, S. T., &Sumowski, J. F. (2012). What variables predict quality of text notes and are text notes related to performance on different types of tests?.*Applied Cognitive Psychology*, *26*(1), 104-117.

[13]. Schaback, R. (2020). On covid-19 modelling. *Jahresbericht der DeutschenMathematiker-Vereinigung*, *122*(3), 167-205.

[14]. Sivamani, S., Binnal, P., Roy, C., Al Khaldi, A., Al Hamar, F., Maran, J. P., ... & Karuppiah, P. (2021). Optimization and characterization of pectin recovered from Persea americana peel using statistical and non-statistical techniques. *Biomass Conversion and Biorefinery*, 1-14.