Advances and Applications in Mathematical Sciences
Volume 20, Issue 10, August 2021, Pages 2313-2331

MODELING AND POTENTIAL PROGNOSIS OF THE NUMBER OF CASES COVID-19 PANDEMIC WITH LINEAR, NONLINEAR REGRESSION MODELS AND ARTIFICIAL NEURAL NETWORK MODELS

VINOTH BALAKRISHNAN, LIJI SEBASTIAN and S. RITA

Assistant Professor
Kristu Jayanti College
Bengaluru, India

Associate Professor
Periyar University
Selam, India
E-mail: vinothb@kristujayanti.com
liji.g@kristujayanti.com

Abstract

Coronavirus disease 2019 (COVID-2019) has been identified as a global threat, and many experiments are being performed using various mathematical models to forecast this epidemic's possible evolution. Many of the biggest wealth Economies are stressed because this disease is highly contagious and transmissible. Because of the rise in number of cases and their resulting burden on the government and health care practitioners, some predictive methods for predicting the number of cases in the future will be needed. We evaluated the performance of the linear, non-linear regression and artificial neural network models to forecast the cases reported daily COVID-19 in India 60 days ahead, and the impact of preventive measures such as social isolation, wearing mask and lockdown on COVID-19 spread. Predicting different parameters (number of positive cases, number of cases reported, number of deaths).

Introduction

Who first discovered coronaviruses?

Avian contagious bronchitis was first reported in newborn chicks in 1931.

2010 Mathematics Subject Classification: 62K20.
Keywords: COVID-19, linear and non-linear regression models, artificial neural network model, R^2, Adjusted R^2, MSE, MAE and RMSE.
Received April 3, 2021; Accepted April 29, 2021

Fred Beaudette and Charles Hudson, from the New Jersey Agricultural Experiment Station, were found to be attributable to a virus in 1937. (J Am Vet Med Ass 1937; 90: 51-8).

Virologist David Tyrrell, Director of the Harnham Down, Joint Cold Research Unit, Wiltshire, published a paper in the British Medical Journal in the year 1965 describing the virus they refer to as B814 and identifying it as a source of common cold. In the 1965s the journal was called B814.They attempted to classify other viruses, but had no luck, and concluded that the viruses they discovered were rhinoviruses. Coronavirus disease (COVID-19) is an infectious disease affected by a coronavirus that has recently been detected. The virus epidemic is believed to have originated in animals and was first spread to humans in the Chinese province of Wuhan in November/December 2019.Artificial neural networks (ANN) have been shown to be a more controlling and self-adaptive method of estimating yield compared to customary linear and simple non-linear analyzes (Simpson [13], Baret et al. 1995, Jiang 2000). In order to learn the dynamic conscious decision between input and output training results, this approach employs a non-linear response function that iterates several times in a particular network structure. For these reasons, the ANN definition was widely used for model creation, especially in highly nonlinear, complicated systems (e.g., Louis and Yan 1998).

Methodology

The data in this study sets involve the number of positive COVID-19 pandemic cases belonging to the between 01/01/2021 and 13/03/2021 in India. These were retrieved by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (http://github.com) accessed on 18/03/2021. In this study, the data is modeled via some curve estimation models to estimate the number of positive COVID-19 positive cases, recovered cases and death cases. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 28 days. The dataset used in the report includes a constant time series overview table that includes the number of reported cases, deaths, and recoveries in the days after the pandemic began.

Some curve estimation models

We consider modeling between the dependent and independent variable. As a result, this approach can be used to determine the relationship between independent and dependent variables, as well as for forecasting.

Linear: $y = b_0 + (b_1 * t)$. Time series values are modeled as a linear function of time.

Quadratic: $y = b_0 + (b_1 * t) + (b_2 * t * 2)$. The quadratic model can be used for the modeling of "taking off" or damping series.

Cubic: $y = b_0 + (b_1 * t) + (b_2 * t * 2) + (b_3 * t * 3)$.

Non-Linear Models

In parametric model, different non-linear models (Bard [1]; Draper and Smith [4]; Montgomery et al. [8]; Ratkowsky [10]; Seber and Wild [12]) given in Table 1 were employed. The model with the most modified R^2 with large F value was chosen from among the nonlinear models in order to achieve a satisfactory fitness test (Montgomery et al. [8]). In case of more than one model being the good fit for the data, the best model has been selected based on lower values of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. Levenberg-Marquardt algorithm (Ratkowsky 1990), in order to achieve detection accuracy, various sets of initial parameter values were sought.

Table 1. Non-linear Models.

Model No.	Model	Name of the Model
I.	$Y = A/(1 + B * EXP\ (-C * X)) + e$	Logistic
II.	$y = e**(b_0 + (b_1 * t))$	Growth
III.	$y = b_0 * (e**(b_1 * t))$	Exponential
IV.	$Y = A * EXP\ ((-(B - X)**2/(2 * C**2)) + e$	Gaussian
V.	$Y = A - B * EXP\ (-C * X**D) + e$	Weibull
VI.	$Y = A * (B**X) * (X**C) + e$	Hoerl
VII.	$Y = A + B * COS\ (C * X + D) + e$	Sinusoidal

VIII.	$Y = C - (C - B)\exp(-A * X) + e$	Monomolecular

Artificial neural networks

An ANN is a biological neural networks-based mathematical or computational model. It comprises an integrated artificial neuron group and processes information by means of a connectionist calculation approach. ANNs are not linear mathematical simulation methods in more realistic words. It can be used for modeling complex input and output relationships, or for finding patterns in data. Mcculloch and Pitts [7] suggested for the first time the concept of a neural artificial network, but they were not very used until the reverse propagation algorithm had been developed because of a lack of computer facilities (Rumelhart et al [11]). Multilayer feeding neural network (MLP) is very common and used for a broad array of purposes rather than other neural network types. Multilayer feeds the neural network learned by back propagation algorithm is built on a controlled method, i.e. the network builds a model based on examples of known data performance. The model must be constructed entirely from given scenarios, which are supposed to provide indirectly the requisite details to create the relationship.

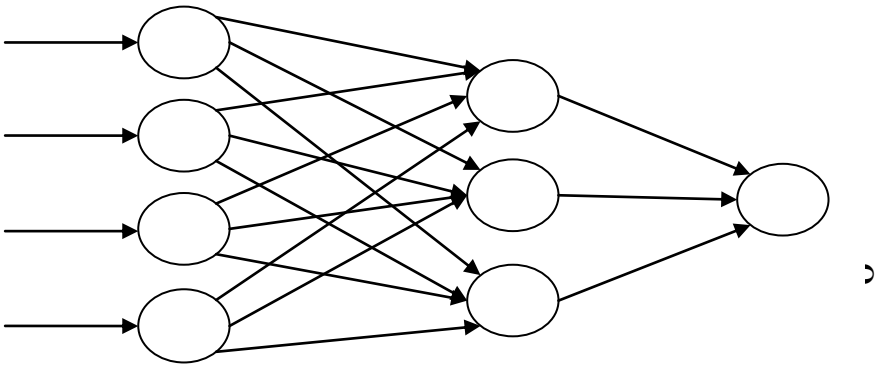

Figure 1. Architecture of neural network (MLP).

An MLP is a strong method that can also model dynamic and variable relations. The output model prediction for a specific input element is permitted. The architecture of the MLP is a layered neural feed network that arranges non-linear components in successive strata and uni-directionally flows information from the input to the output strata through hidden layers.

Back Propagation Algorithm

Back propagation is a popular method of training artificial neural networks to reduce the objective function to the smallest possible value. The Multilayer Perceptron network is formed by means of one of the supervised learning algorithms, which uses the data to change the weights and thresholds of the network to reduce error in the predictions of the training set. The most popular example is back propagation. First, it computes the total weighted input x_j, using the formula:

$$X_j = \sum_i y_i W_{ij}.$$

Typically we use the hyperbolic tangent function:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

The backpropagation algorithm consists of four steps:

(i). Compute how fast the error changes as the activity of an output unit is changed. This error derivative (EA) is the difference between the actual and the desired activity.

$$EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j.$$

(ii). Compute how fast the error changes as the total input received by an output unit is changed. This quantity (EI) is the answer from step (i) multiplied by the rate at which the output of a unit changes as its total input is changed.

$$EI_j = \frac{\partial E}{\partial X_j} = \frac{\partial E}{\partial y_j} \times \frac{\partial y_j}{\partial x_j} = EA_j y_j (1 - y_j).$$

(iii). Compute how fast the error changes as the weight on the connection into an output unit is changed. This quantity (EW) is the answer from step (ii) multiplied by the activity level of the unit from which the connection emanates.

$$EW_{ij} = \frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial X_j} \times \frac{\partial X_j}{\partial W_{ij}} = EI_j y_i.$$

(iv). Compute how fast the error changes as the activity of a unit in the previous layer is changed. When the function of a node in the activation function changes, the activity of all the output layer to which it is related changes. We apply all these different impacts on production units to measure the total influence on the error. It is the answer in step (iii) multiplied by the weight on the connection to that output unit.

$$EA_i = \frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial x_j} \times \frac{\partial x_j}{\partial y_i} = \sum_j EI_j W_{ij}.$$

We can translate the EAs of one layer of units into EAs for the previous layer by using steps (ii) and (iv). This method can be replicated as many times as required to obtain EAs for previous layers. Once we know the EA of a machine, we can calculate the EWs on its connection requests using steps (ii) and (iii).

Measures of goodness of fit:

In this study, we measure the performance of each of the models in terms of R-squared (R^2) score, Adjusted R-squared (R^2_{adjusted}), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE).

***R*-Squared and Adjusted *R*-Squares Score:**

To test the goodness of fit of the fitted polynomial model, the coefficient of determination defined as the proportion of total variation in the response variable (time) being explained by the fitted model is widely used. The high R^2 score shows the goodness of the fitted model. R^2 is a linear model that explains the percentage of variation independent variable. It can be found as:

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

The adjusted R-squared (R^2_{adjusted}) is a modified form of R^2, which also like R^2 shows how well the data points fit for the curve. The main distinction between R^2 and R^2_{adjusted} is that the latter takes into account the number of features in a prediction model. The R^2_{adjusted} can be defined as:

$$R^2_{\text{adjusted}} = 1 - (1 - R^2)\frac{n - 1}{(n - k)}.$$

Here, n is the sample size and k is the number of independent variables in the regression equation.

Measures of the Adequacy of the fitted Model

In addition to the above, three more reliability statistics viz., Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) are generally utilized to measure the adequacy of the fitted model and it can be computed as follows:

The mean absolute error is the average magnitude of the errors in the set of model predictions.

$$MAE = \sum_{i=1}^{n} \left| (Y_i - \hat{Y}_i) \right| / n.$$

Another approach to test the efficiency of regression models is to use mean square error. MSE measures the difference between data points and the regression line by squaring them. The lower the mean squared error, the closer you are to choosing the best fit line. It can be defined as

$$MSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 / n.$$

Root mean square error can be defined as the standard deviation of the prediction errors. Prediction errors also known as residuals are the distance from the best fit line and actual data points. It is the error rate given by the square root of MSE given as:

$$RMSE = \left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 / n\right]^{1/2}.$$

The lower the values of these statistics, the better are the fitted model.

Results and Discussion

Some parametric and non-parametric models have been used to model the number of cases, number of recovered and number of deaths suffering from the COVID-19 epidemic depending on the days in India. This research aims to estimate the number of people who could be affected in terms of new infected cases and deaths, as well as the number of predicted recoveries over the next 28 days. Six models Linear, Quadratic, Growth, Exponential, Logistic and Artificial Neural Network (Multilayer Perceptron Model) have been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries.

A. New Infected Confirm Cases Future Forecasting

The new confirmed cases of COVID-19 increase day by day Table 1 shows the forecasting results of the models used in this study. Cubic and quadratic lead the table in terms of performance, linear, growth, logistic also performed good, while exponential and ANN-MLP performs very poorly in terms of all the evaluation criteria. Graphs in figures 2, 3, 4, 5, 6 and 7 show the predictions of curve estimation models.

Table 2. Models performances on future forecasting for new infected confirm cases.

Model	R^2	Adjusted R^2	MAE	MSE	RMSE
Linear	0.99	0.99	16258.79	366168176	19135.52
Quadratic	0.99	0.99	16358.16	362828131	19048.04
Cubic	1.00	1.00	2820.98	14158348	3762.75
Growth	0.99	0.99	16436.68	364086997	19081.06
Exponential	0.99	0.99	199713.65	55739558093.87	236092.27
Logistic	0.99	0.99	16436.68	364086997	19081.06
ANN-MLP	0.99	0.99	20597.83	838419522	28955.47

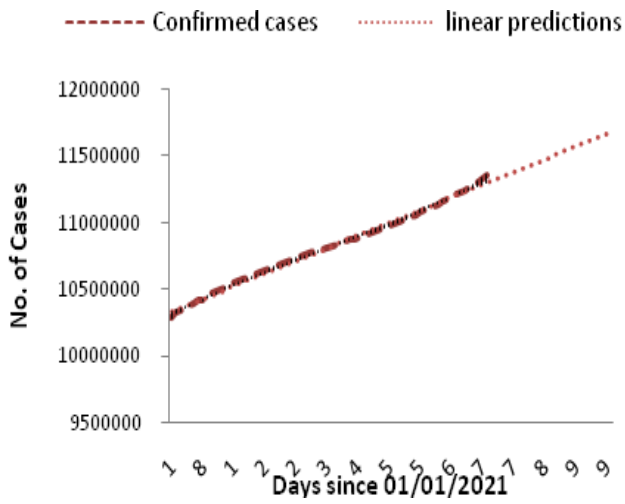

Figure 2. New infected confirm cases prediction by Linear for the upcoming 28 days.

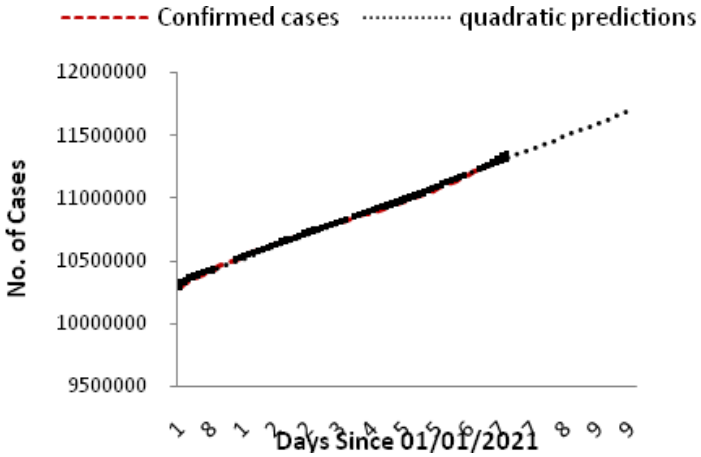

Figure 3. New infected confirm cases prediction by Quadratic for the upcoming 28 days.

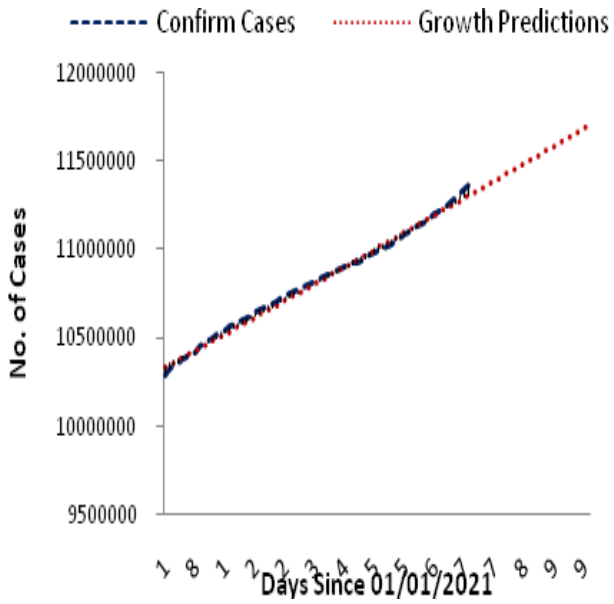

Figure 4. New infected confirm cases prediction by Growth for the upcoming 28 days.

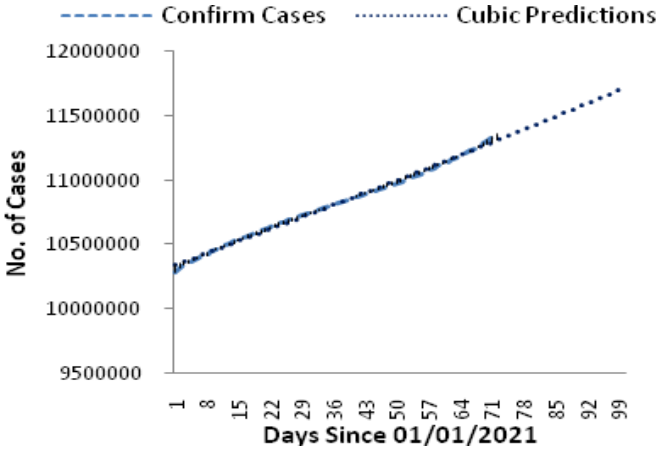

Figure 5. New infected confirm cases prediction by Cubic for the upcoming 28 days.

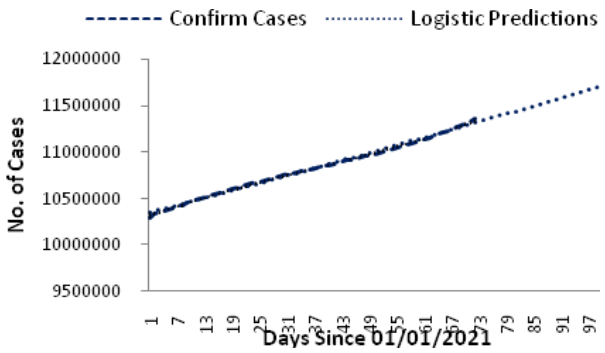

Figure 6. New infected confirm cases prediction by Logistic for the upcoming 28 days.

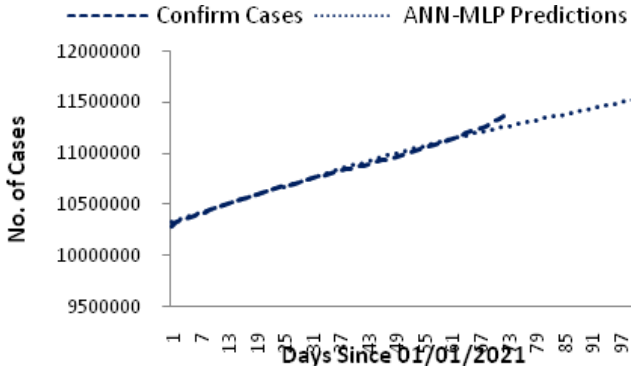

Figure 7. New infected confirm cases prediction by ANN-MLP for the upcoming 28 days.

B. Recovery Rate Future Forecasting

In recovery future forecasting the Cubic again performs better among all the other models. All other models perform poorly, the order of performance from best to worst is cubic followed by quadratic, linear and growth, exponential, and logistic has same evaluation metrics and ANN-MLP perform very poorly compare than above mentioned models. The prediction trends for the coming days are shown in figures 8, 9, 10, 11, 12, 13 and 14 show the predictions of leaning models. The performance results of learning models are shown in the Table 3 below.

Table 3. Models performances on future forecasting for recovery rate.

Model	R^2	Adjusted R^2	MAE	MSE	RMSE
Linear	0.99	0.99	21555	672677383	25936.02
Quadratic	0.99	0.99	10720	184668924	13589.29
Cubic	1.00	1.00	5254	37449653	6119.61
Growth	0.99	0.99	24705	856303797	29262.69
Exponential	0.99	0.99	24705	856303797	29262.669
Logistic	0.99	0.99	24705	856303797	29262.669
ANN-MLP	0.99	0.99	17540.44	509671311.75	22575.90

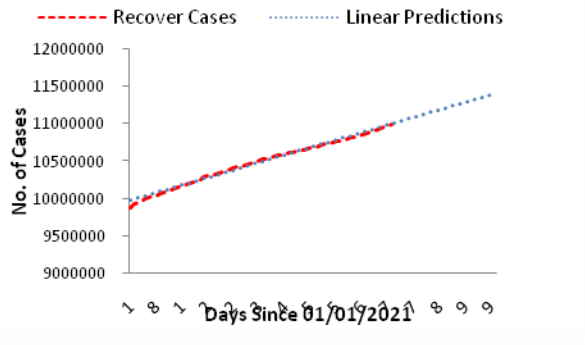

Figure 8. Recovery rate prediction by Linear for the upcoming 28 days.

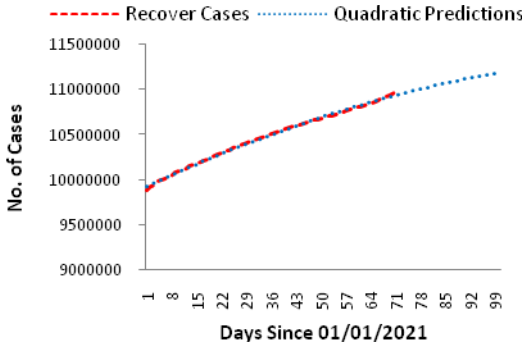

Figure 9. Recovery rate prediction by Quadratic for the upcoming 28 days.

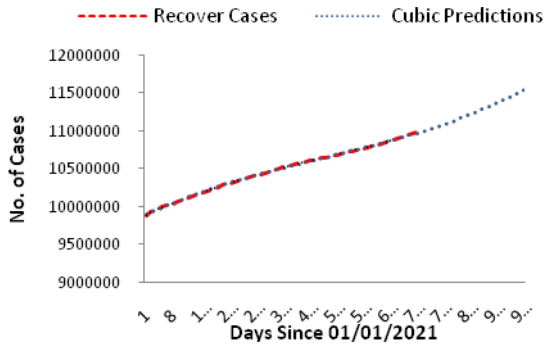

Figure 10. Recovery rate prediction by Cubic for the upcoming 28 days.

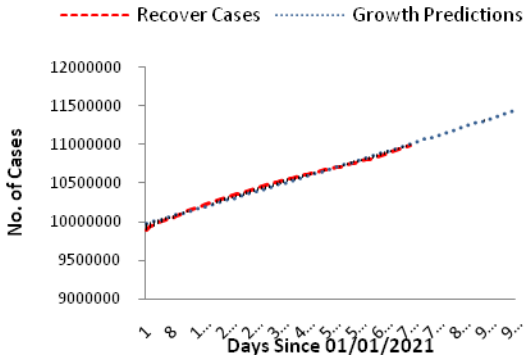

Figure 11. Recovery rate prediction by Growth for the upcoming 28 days.

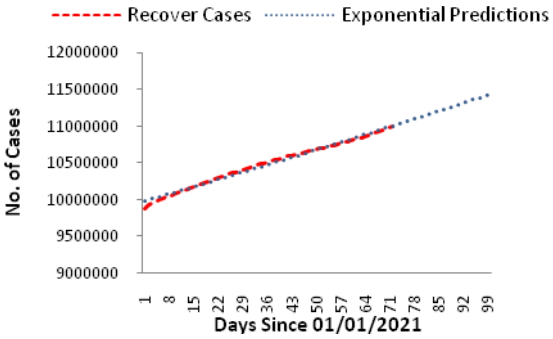

Figure 12. Recovery rate prediction by Exponential for the upcoming 28 days.

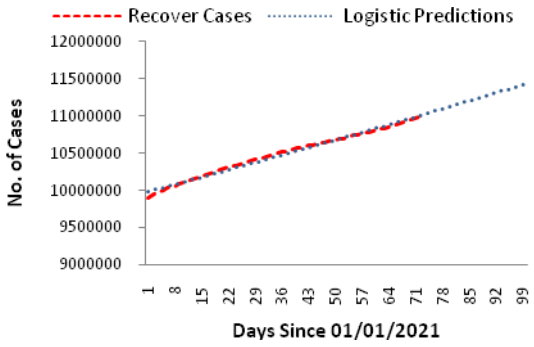

Figure 13. Recovery rate prediction by Logistic for the upcoming 28 days.

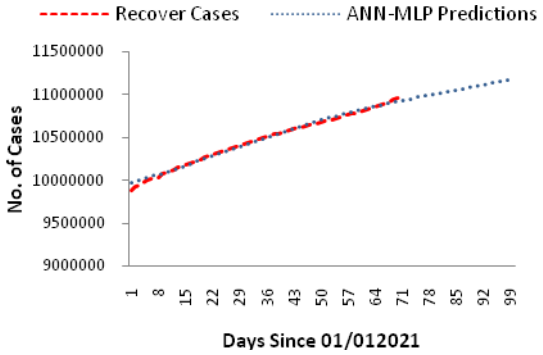

Figure 14. Recovery rate prediction by ANN-MLP for the upcoming 28 days.

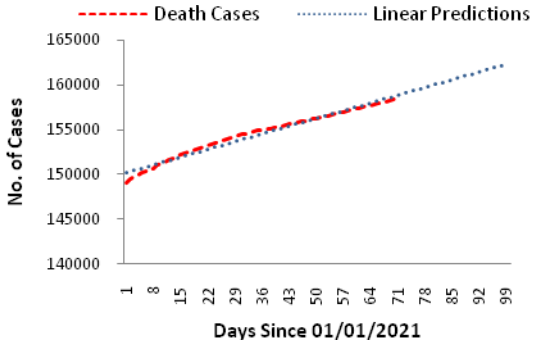

Figure 15. Death prediction by Linear for the upcoming 28 days.

C. Death Rate Future Forecasting

The study performs predictions on death rate and according to results

cubic performs better among all the models, quadratic and ANN-MLP performs equally well and achieves almost the same R2 value. In comparison, growth, exponential, logistic, and linear worst in this situation. The results are shown in Table 4.

Table 4. Models performance on future forecasting for death rate.

Model	R^2	Adjusted R^2	MAE	MSE	RMSE
Linear	0.98	0.98	307.48	141370.77	375.99
Quadratic	0.99	0.99	153.30	31439.14	177.31
Cubic	1.00	1.00	30.47	1548.97	39.35
Growth	0.98	0.98	323.49	154083.2	392.53
Exponential	0.98	0.98	323.49	154083.2	392.53
Logistic	0.98	0.98	323.49	154083.2	392.53
ANN-MLP	0.99	0.99	193.63	54329.53	233.08

Figures 15, 16, 17, 18, 19, 20, and 21 shows the performance of linear, quadratic, cubic, growth, exponential, logistic and ANN-MLP models respectively in the form of graphs.

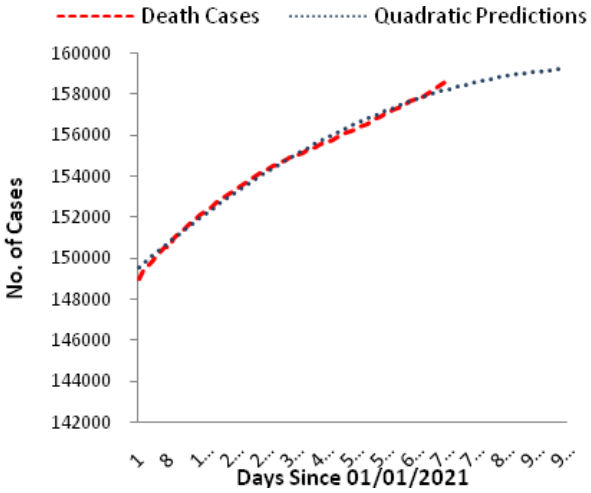

Figure 16. Death prediction by Quadratic for the upcoming 28 day.

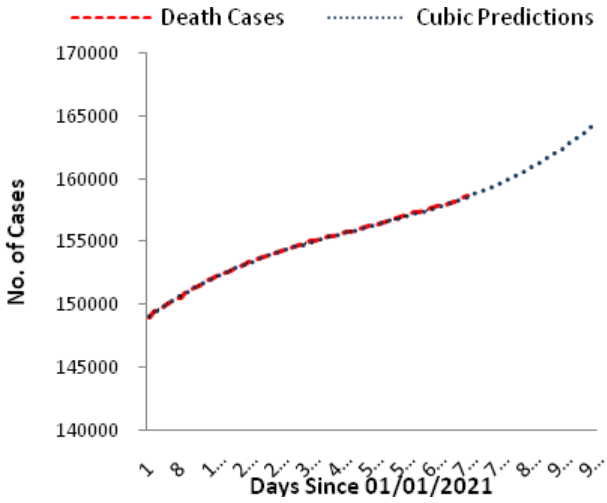

Figure 17. Death prediction by Cubic for the upcoming 28 days.

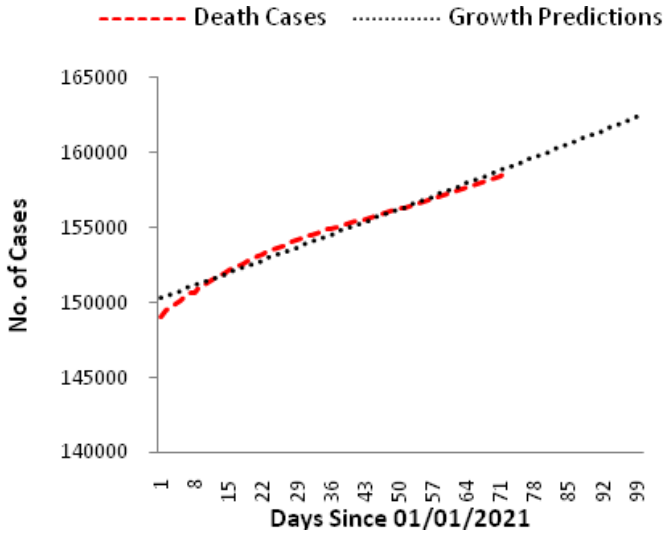

Figure 18. Death prediction by Growth for the upcoming 28 days.

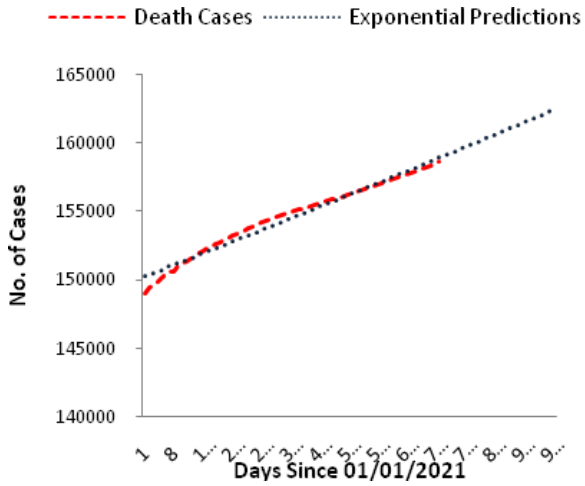

Figure 19. Death prediction by Exponential for the upcoming 28 days.

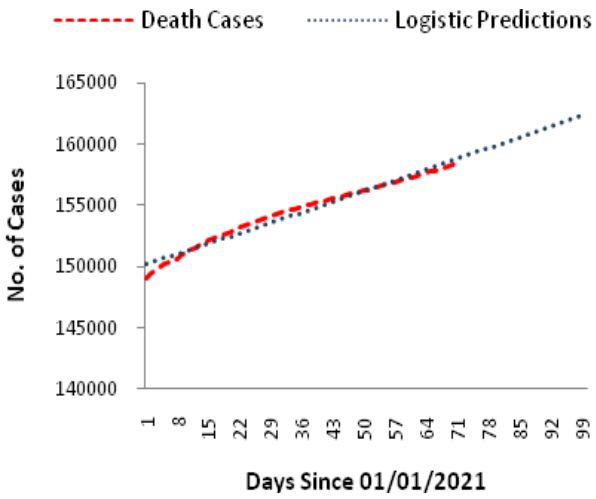

Figure 20. Death prediction by Logistic for the upcoming 28 days.

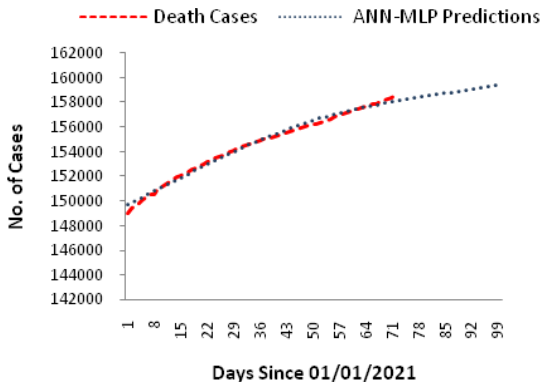

Figure 21. Death prediction by ANN-MLP for the upcoming 28 days.

Graphs in all figures predict that the death rate will increase in upcoming days which is a very alarming sign.

Conclusion

In this paper, the curve estimation methods- linear methods, non-linear regression methods/ANN-MLP method has been used to estimate the possible number of positive cases of COVID-19 in India for the next 28 days. The number of confirm cases, recovered cases and death rate has also been estimated by using Cubic estimation methods. Cubic curve estimation methods perform well for forecasting to some extent to predict confirm cases, recovered cases and death rate. According to the results of this model, the death rate will increase in upcoming days, and recoveries rate also will increase slowly. Overall we conclude that model predictions according to the current scenario are correct which may be helpful to understand the upcoming situation in India. The effect of preventing measures like social isolation, lock down, wearing mask and using sanitizer has also been observed which shows that by these preventive measures, spread of the virus can be reduced significantly.

References

[1] Y. Bard, Nonlinear Parameter Estimation, Academic Press: New York 1974.

[2] F. R. Beaudette and B. D. Hudson, Cultivation of the virus of infectious bronchitis, Journal of the American Veterinary Medical Association 90(1) (1937), 51-60.

[3] N. R. Draper and H. Smith, Applied Regression Analysis, 3rd Edition, John and Wiley and Sons, USA 1998.

[4] R. Farebrother, Further results on the mean square error of ridge regression, Journal of the Royal Statistical Society Series B (Methodological) 38(3) (1976), 248-250.

[5] Furgan Rustan et al., COVID-19 future forecasting using supervised machine learning models, IEEE Access 8 (2020), 101489-101498

[6] Johns Hopkins CSSE, Novel Coronavirus (COVID-19) Cases, 2020, https://github.com/CSSEGISandData/COVID-19

[7] W. S. McCulloch and W. Pitts, A Logical Calculus of Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics (1943) 115-133.

[8] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis, John Wiley and Sons, Inc 2003.

[9] C. R. Rao and H. Toutenburg, Linear models, Linear Models; Springer (1995), 3-18.

[10] D. A. Ratkowsky Handbook of Non-linear Regression Models, Marcel Dekker, New York (1990).

[11] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating error, Nature 323 (1986), 533-536, Reprinted in Anderson and Rosenfeld (1988), 696-699.

[12] G. A. F. Seber and C. J. Wild, Non-Linear Regression, John Wiley and Sons, New York (1989).

[13] P. K. Simpson, Foundations of neural networks, United States: N. P., 1994. Web.