 doi:10.20944/preprints202008.0446.v1

Modeling and Prediction of COVID-19 Outbreak in India

Saurabh Kumar[1], Varun Agiwal[2], Ashok Kumar[3] and Jitendra Kumar[4,*]

Abstract: As the outbreak of coronavirus disease 2019 (COVID-19) is continuously increasing in India, so epidemiological modeling of COVID-19 data is urgently required for administrative strategies. Time series and is capable to predict future observations by modeling the data based on past and present data. Here, we have modeled the epidemiological COVID-19 Indian data using various models. Based on the collected COVID-19 outbreak data, we try to find the propagation rule of this outbreak disease and predict the outbreak situations in India. For India data, the time series model gives the best results in the form of predication as compared to other models for all variables of COVID-19. For new cases, new deaths, total cases and total deaths, the best fitted ARIMA models are as follows: ARIMA(0,2,3), ARIMA(0,1,1), ARIMA(0,2,0) and ARIMA(0,2,1). Based on time series analysis, we predict all variables for next month and conclude that the predictive value of Indian COVID-19 data of total cases is more than 20 lakhs with more than 43 thousand total deaths. The present chapter recommended that a comparison between various predictive models provide the accurate and better forecast value of the COVID-19 outbreak for all study variables.

Keywords: COVID-19, Time series model, Indian data, Exponential model and 3rd- degree polynomial model

[1] S. Kumar, *Department of Management, Invertis University, Bareilly, Uttar Pradesh, India.*
e-mail: *sonysaurabh123@gmail.com*

[2] V. Agiwal, *Department of Community Medicine, Jawaharlal Nehru Medical College, Ajmer, Rajasthan, India.*
e-mail: varunagiwal.stats@gmail.com

[3] A. Kumar, *Department of Community Medicine, SHKM Government Medical College, Nuh, Haryana, India.*
e-mail: ashokkr.166@gmail.com

[4, *] J. Kumar (Corresponding Author), *Department of Statistics, Central University of Rajasthan, Ajmer, Rajasthan, India*
e-mail: vjitendrav@gmail.com

 doi:10.20944/preprints202008.0446.v1

1 Introduction

The novel coronavirus COVID-19 is a pandemic disease that has an unprecedented challenge for the health of human life worldwide. This disease is spreading from person to person contact during speaking or sneezes, touching surfaces, and use other surrounded virus objects. Since the starting of the COVID-19 disease, the proportion of the affected world population and total infected land area is rapidly increased day by day because no vaccines and antiviral drugs are present to control the infections. So, the outbreak of COVID-19 to be controlled using proper planning and policies taken by the local government as well as followed instructions provided by the World Health Organization (WHO) from time to time.

COVID-19 first case was declared on the 20th of January 2020 in India and after that cases are progressively increasing day by day. So, the government imposed an effective lockdown across the country from 23 March 2020 to control the spread of infection at the community level. The strategy has resulted that the growth of coronavirus cases is not so much higher as recorded at the global level. In current times, India has observed more than 10,00,000 cases with more than 63% recovered cases, an approximate mortality rate of 2.5% which is low as compared to the global level (7%). Since the declaration of the present pandemic, many research papers are published on various aspects of COVID-19. Maleki *et al.* (2020) used the autoregressive model to forecast the "confirmed" and "recovered" COVID-19 worldwide based on two-piece scale mixture normal distribution and evaluated the performance with standard Gaussian autoregressive model. The results indicated that the proposed method performed well in forecasting confirmed and recovered COVID- 19 cases in the world. Qi *et al.* (2020) examined the associations of daily average temperature and relative humidity with the daily count of COVID-19 cases in 30 Chinese provinces using a generalized additive time series model. Jiang *et al.* (2020) found the disease transmission of the COVID-19 outbreak based on a dynamic model, time-series approach, and data mining technique. Then, they predicted the epidemic situation under the best suitable model and proposed an effective control and prevention method. Sujath *et al.* (2020) performed linear regression, multilayer perceptron, and vector autoregressive models on the Kaggle dataset for anticipating the future effects of COVID-19 pandemic in India. Pandey and Samanta (2020) considered the autoregressive integrated moving average model to forecast the daily COVID-19 cases of India whereas Tomar and Gupta (2020) predicted the COVID-19 cases based on long short-term memory method. Both predictive models are less or more useful to show the future trend of the COVID-19 outbreak. Salgotra *et al.* (2020) presented prediction models based on genetic programming for times series prediction of confirmed and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0446.v1

death cases of the COVID-19 in whole India and provided the study for three most affected states namely Maharashtra, Gujarat, and Delhi.

In the present chapter, time series and non-linear regression models are studied to know the better future trend of the COVID-19 variables. For this, an autoregressive integrated moving average (ARIMA) model has been used to fit the data and analyze accordingly. Since the increase in the numbers of cases, this data is increasing exponentially with the non-linear trend so non-linear models are also useful to analyze the same data. Here, we have considered the 3rd-degree polynomial regression and the exponential model for fitting the data and obtain the optimal model that makes the prediction better. The main contribution of the present is to know the present trend of COVID-19 cases and achieve the most favorable model that can efficiently predict and estimate the confirmed and death COVID-19 cases in India.

2 Methods

2.1 Study design and Indian data sources

This is a need of data source where all states as well as national level Indian COVID-19 data are recorded on daily as well as cumulative bases. For our analysis, the data is downloaded from www.covid19india.org. This is an online platform which is providing the real-time coverage of COVID-19 outbreak in India and state wise, obtain by state press bulletins and official handles documents like twitter, report, etc. Every source of information linked to provide for more detailed information on individual and number of cases and published on API (https://api.covid19india.org/). We closely observe the updated data on COVID-19 India between 14th March 2020 to 13th July 2020 to record the cumulative number of cases, the cumulative number of deaths, new cases, and new deaths. The whole dataset splits into three sets: training, testing, and forecast.

- Training date: 14th March 2020 to 30th June 2020
- Testing data sets to test the accuracy of prediction: 1st July 2020 to 13th July 2020.
- Forecast period: 14th July 2020 to 22nd August 2020.

2.2 Statistical modeling

Based on the collected COVID-19 outbreak data, we try to find the propagation rule of this outbreak diseases and predict the outbreak situations in India. There are generally three kinds of methods like a dynamic model of infectious, data mining technology, and statistical modeling to study the law of infectious disease transmission. In this study, we use statistical modeling based on a random process, ARIMA time series model, and other statistical models such as 3rd- degree polynomial and exponential trend model.

The autoregressive integrated moving average (ARIMA) model is generally used to describe the stochastic process that varies over time. This model is applied in those cases where series show non-stationarity. They could have non-constant mean, time-varying second moment such as non-constant variance or both. The initial difference or transform step can be applied one or more than one time to eliminate the non-stationarity effect. The general mathematical form of the ARIMA(p,d,q) model with usual notations is as follows;

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\omega_t \qquad (1.1)$$

where ϕ and θ are the autoregressive and moving average lag coefficients, B is a backward operator. For modeling the infected and the deaths cases due to COVID-19, the model has been fitted with the best order of p, d, and q with the help of various selection procedures. This model has a very proven application in various areas from economics to health science. Researchers do the modeling based on the ARIMA model and make the prediction. Song *et al.* (2016) predicted the monthly incidence of influenza in China for 2012 using the ARIMA model. Yin (2020) proposed a time-series prediction model for the mutation prediction of influenza A viruses. Based on the time series model, Maleki *et al.* (2020), Chimmula and Zhang (2020), Yonar *et al.* (2020), and Papastefanopoulos *et al.* (2020) studied the countries-wise and worldwide COVID-19 cases.

If the structure of the series is seen to be curvilinear then the best-fitted model is the polynomial regression with nth degree function. In this regression, the relationship between dependent and independent variables is modeled such as the dependent variable is an nth degree function of an independent variable. Here, we use the least-square cubic or 3rd-degree polynomial model for our analysis. The form of the model is;

3rd-degree Polynomial: $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \omega_t$ (1.2)

Another model is an exponential trend model where series begins slowly and then increased rapidly to achieve the peak and after then the growth slows down at a maximum rate. The mathematical form of this model is;

Exponential Model: $\log(y_t) = \log(a) + \log(b)\,t + \omega_t$ (1.3) where, ω_t is the error terms from the white noise process. These non-linear models are well discussed by Makridakis *et al.* (1982), Taylor (2003), Chatterjee and Sarkar (2009), Zhang et al. (2014), Ma and Zheng (2017), Shastri *et al.* (2018) and Yadav (2020) at a different aspect of situations special for disease analysis. The best model is selected by using the Akaike Information Criterion (AIC), coefficient of determination (R^2) and root mean square error (RMSE). The correlogram and Q-Q plot are drawn for the residuals of the fitted model. These two plots are used to test whether the residuals having any serious dependency or not and it follows the white noise process or not. The forecast performance is judged by root means square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MPE). Using these accuracy matrices, we evaluate the forecast irrespective of the given scale of the series.

3 Results

In this section, we record the results of the fitted model for the COVID-19 data. This data set consists of new cases, new deaths, total cases, and total deaths for COVID-19. The data are inputted in Excel 2016 and R software (version 3.6.3; 29 February 2020). First, we compute the average and standard deviation of the selected COVID-19 variables and recorded in Table 1.1. The fluctuation is more in new cases and total cases as compared to new deaths and total deaths. This value indicates that there is more variability between the new conformed cases and the number of affected cases due to COVID 19 is very rapid in India. The stationarity of the series is accessed through the ADF test and results are also shown in Table 1.1. The p-value of the ADF test is greater than the level of significance i.e., 0.05, it means that all four selected COVID-19 variables are non-stationary with the trend.

Table 1.1: Summary of four variables and stationary test of COVID-19 series in India

Variable	Mean	SD	Augmented Dickey-Fuller (ADF)	
			Statistics	P-value
New Cases	5423	5592.198	-0.321	0.988
New Deaths	161	226.727	-3.537	0.052
Total Cases	134380	162588.052	4.681	0.990
Total Deaths	4083	4972.983	1.870	0.990

The deterministic information is extracted from the original time series and concludes that three variables (New Cases, Total Cases, Total Deaths) are stationary at second-order difference whereas New Deaths is stationary at one order difference. The differential time series is verified as non-random series through a white noise test. The cubic and exponential models are also fitted and corresponding AIC, R^2 and RMSE is recorded for all four variables in Table 1.2.

Table 1.2: Summary of the fitted model of COVID -19 variables in India

Variable	Model	AIC	R^2	RMSE
New Cases	ARIMA(0,2,3)	**1685.118**	**0.987**	**641.732**
	Polynomial Regression (k=3)	2141.358	0.986	7764.885
	Exponential Regression	2039.621	0.891	4940.134
New Deaths	ARIMA(0,1,1)	1427.678	0.452	279.041
	Polynomial Regression (k=3)	**1417.423**	**0.497**	**272.003**
Total Cases	ARIMA(0,2,0)	**1695.222**	**1.000**	**705.293**
	Polynomial Regression (k=3)	2853.973	1.000	210346.355
	Exponential Regression	2837.500	0.922	198596.197
Total Deaths	ARIMA(0,2,1)	**1397.572**	**0.999**	**170.599**
	Polynomial Regression (k=3)	2099.961	0.996	6410.637
	Exponential Regression	2070.137	0.894	5689.768

Remark: In the case of New death, the Exponential model doesn't fit because of the value of log at zero is infinity.

From Table 1.2, one may observe that the ARIMA time series model recorded better among the selected models for all COVID-19 variables except for new deaths series. In the modeling of new death variable, cubic regression model has been obtained the best fitted model because of the smaller value of AIC and RMSE. The performance of the dynamic time series model is much better as compared to the regression trend model as recorded AIC and RMSE minimum and R^2 value is high as compare to the other two models for all selected variables.

Table 1.3: χ^2 value and p-value of the Box-Ljung test at three different lag values.

Variable	Model	$Lag = 1$		$Lag = 2$		$Lag = 3$	
		Q-value	p-value	Q-value	p-value	Q-value	p-value
New Cases	ARIMA(0,2,3)	0.001	0.972383	1.392	0.707324	12.020	0.034516
New Deaths	ARIMA(0,1,1)	0.001	0.977860	0.248	0.969528	0.748	0.980209
Total Cases	ARIMA(0,2,0)	1.034	0.309320	6.412	0.093196	20.078	0.001208
Total Deaths	ARIMA(0,2,1)	0.197	0.657207	1.255	0.739745	1.671	0.892545

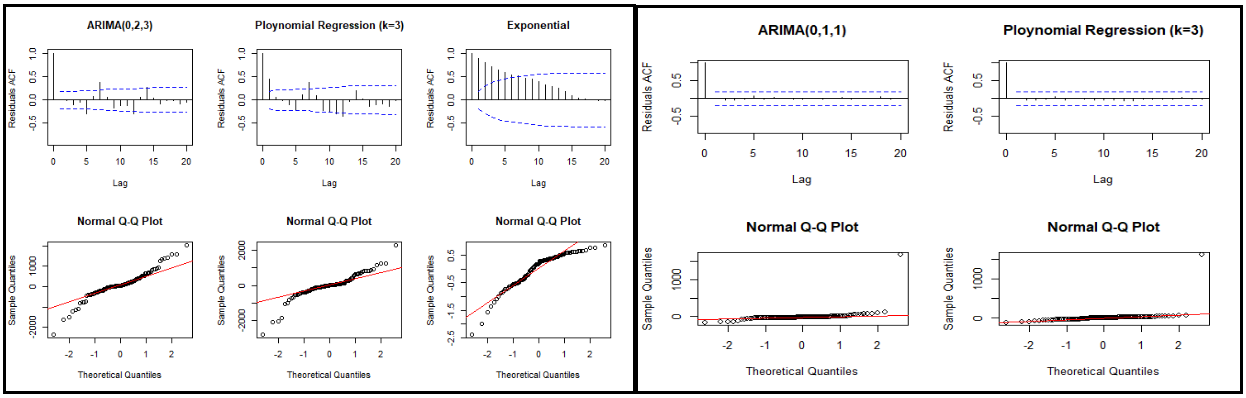
In the above Table 1.3, one can observe that the value of the Box-Ljung p-value decreases as the value of lag increases in the ARIMA residuals. The p-value of this test is greater than the level of significance for new and total deaths i.e., 0.05 whereas the other two variables (new and total cases) p-value is less than 0.05 at lag 3. This indicates that the residuals of the ARIMA model are independently for the cases variable of the COVID-19. For the regression model, we perform the ANOVA test and record the p-value in Table 1.4.

Table 1.4: ANOVA results of fitted polynomial regression and exponential model.

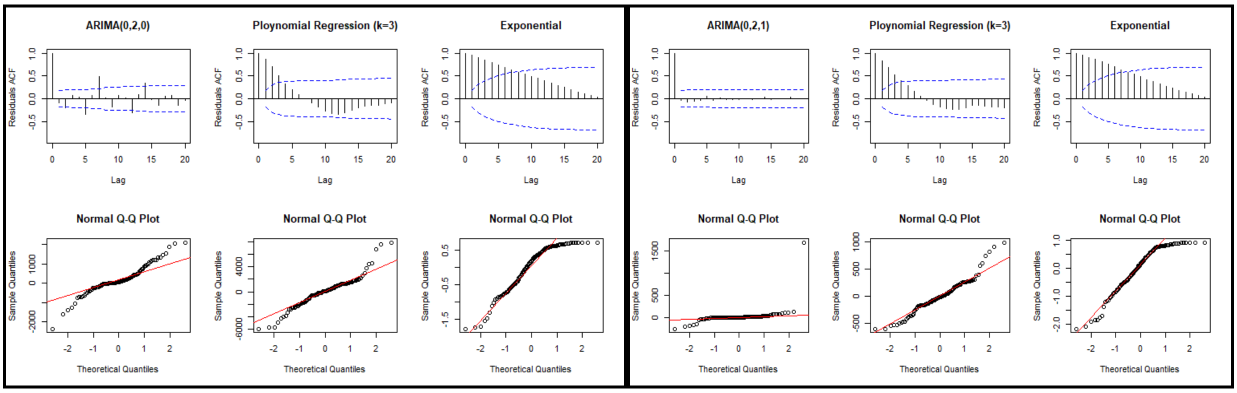
Variable	Model	F	p-value
New Cases	Polynomial Regression (k=3)	2494.6	0.000***
	Exponential Regression	868.93	0.000***
New Deaths	Polynomial Regression (k=3)	34.226	0.000***
Total Cases	Polynomial Regression (k=3)	166122	0.000***
	Exponential Regression	1250.9	0.000***
Total Deaths	Polynomial Regression (k=3)	9393.5	0.000***
	Exponential Regression	890.46	0.000***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In Table 1.4, polynomial and exponential model results show significantly for all four variables of the COVID-19 series. F values represent an improvement in the prediction of the number of COVID cases by the fitting of these models. In Table 1.4, the p-value of all variables shows significant evidence in vouchsafing the model. This indicates that the model is well fitted. The ACF and QQ plot is also plotted for all fitted residual variables in Fig. 1.1.



(a) New Cases (b) New Deaths



(c) Total Cases (d) Total Deaths

Fig. 1.1 Analyzing the residuals from the selected model for COVID -19 series

Fig. 1.1 shows that the ACF plot of new and total cases of ARIMA residuals shows a spine, but it is not quite enough for Box-Ljungs to be significant at the 5% level whereas in the case of the other two variables shows significant autocorrelation with lag. The autocorrelation is not particularly large for ARIMA model up to 5th lag, so it is unlikely to any noticeable impact on the forecasts or the prediction interval of the COVID-19 series at lag 5.

Based on the results provided in Table 1.1, the fitted values and original values are plotted for all four variables in Fig. 1.2. From Fig. 1.2, we observe that the ARIMA and cubic models are very closely fitted to all four observed COVID–19 series. The exponential model was not fitted in New Deaths variable because someday the number of deaths is zero at the beginning and we know that the value of log at zero is infinite. In the rest of the three variables, overall fitting to the exponential model was not up to marks for the other three variables. So, we conclude that the best-fitted model is a time series model that makes the forecast better.

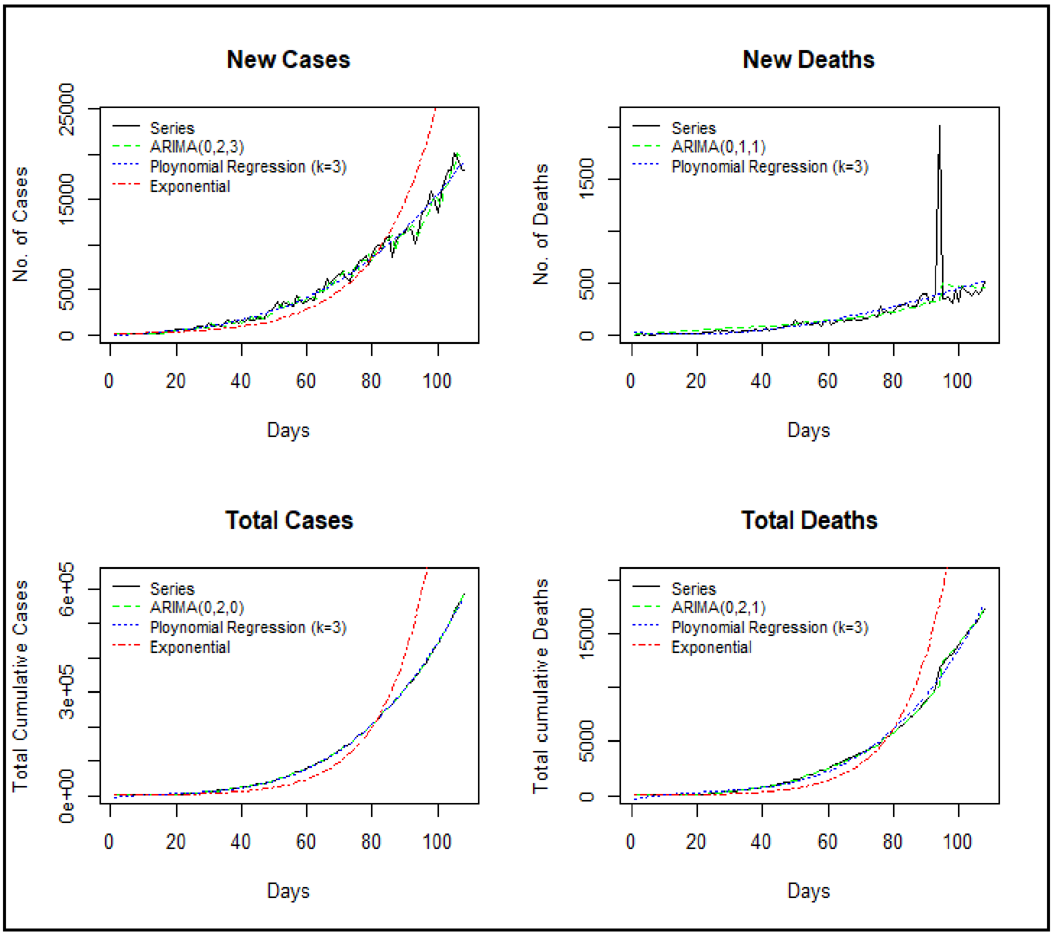


Fig. 1.2 Observed and fitted a series of all variables of COVID -19 based on various models.

After the fitting of the model on all four variables, the model is tested and validated, forecast the value using the best-fitted model. The forecast value is predicting for the period 1st July to 13th July 2020. The forecast graph is shown in Fig. 1.3 for each variable individually.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0446.v1

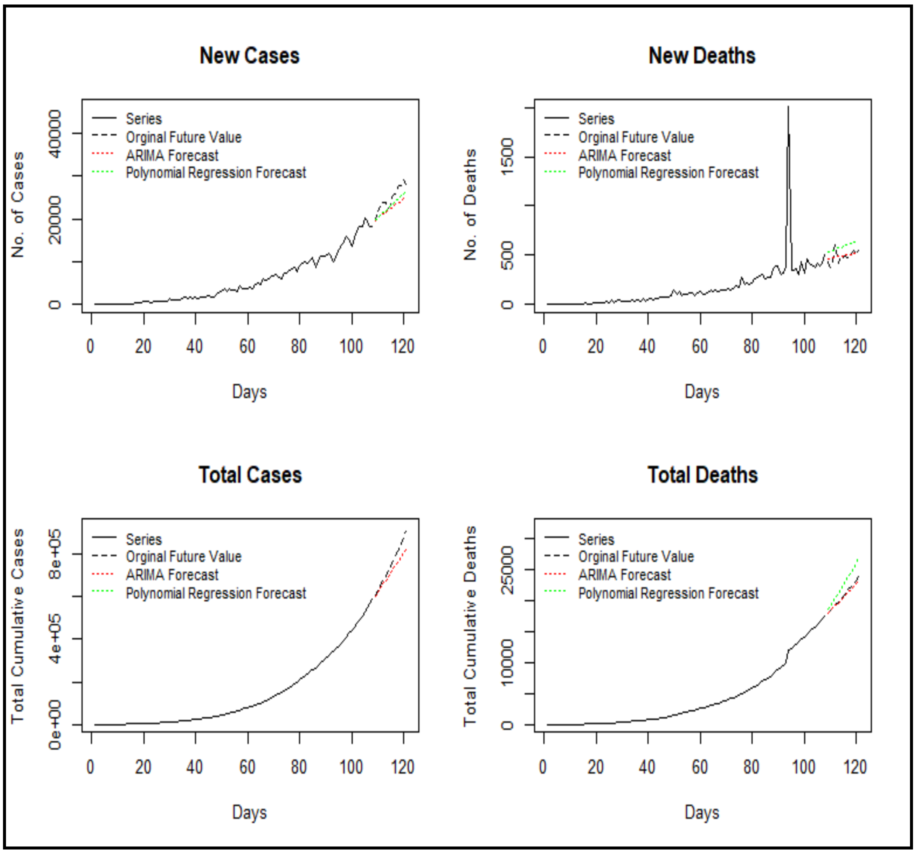


Fig. 1.3 Predication chart of all four variables of COVID -19 based on ARIMA and Polynomial Regression model.

In the forecast plot, obtained the future values based on ARIMA and cubic model only because the predicated value using exponential trend model is not closed to actual values. All fitted models are used to predict the upcoming thirteen days of COVID-19 cases and compared with original data recorded during this period for variables. To compare the forecasted series, we obtained the RMSE, MAE, and MAPE and recorded in Table 1.5.

Table 1.5: Evaluate the forecast accuracy for all variables based on all fitted models

Variable	Model	RMSE	MAE	MAPE
New Cases	ARIMA(0,2,3)	**2577.100**	**2249.582**	**8.679**
	Ploynomial Regression (k=3)	2978.608	2721.682	8.727
	Exponential Regression	37224.091	35856.047	142.603
New Deaths	ARIMA(0,1,1)	**51.338**	**37.195**	**7.798**
	Polynomial Regression (k=3)	116.279	112.910	24.064
Total Cases	ARIMA(0,2,0)	**34198.974**	**31787.692**	**4.021**
	Polynomial Regression (k=3)	45306.833	32766.449	4.448
	Exponential Regression	1907421.022	1813541.959	235.974
Total Deaths	ARIMA(0,2,1)	**365.918**	**289.298**	**1.315**
	Polynomial Regression (k=3)	1854.272	1711.446	8.028
	Exponential Regression	72788.602	68992.257	326.087

In Table 1.5, ARIMA model gives more accurate forecast value as compared to the other two models because the value of RMSE, MAE, and MAPE is very low when we predict it using the ARIMA model. So ARIMA(0,2,3), ARIMA(0,1,1), ARIMA(0,2,0), and ARIMA(0,2,1) time series models are better to use for predicting the possible number of COVID-19 cases occurred in term of new cases, new deaths, total cases and total deaths in India. Based on the best fitted ARIMA model, predict the Covid-19 outbreak in India for the next month. The predicted values are listed for respective variables with the corresponding dates in Table 1.6. We also reported lower and upper confidence limits of the predicted value of the cases at a 95% confidence interval.

Table 1.6: The predicted value of the number of COVID-19 cases till 22nd August 2020 for India scenario

Date	New Cases			New Deaths			Total Cases			Total Deaths		
	Cases	Lower	Upper	Cases	Lower	Upper	Cases	Lower	Upper	Cases	Lower	Upper
14-07-2020	28890	27508	30273	524	209	839	935824	934230	937418	24224	23902	24547
17-07-2020	31498	29681	33315	538	219	857	1020361	1011632	1029090	25712	24905	26520
20-07-2020	33578	31394	35761	552	229	874	1104898	1086041	1123755	27201	25911	28490
23-07-2020	35657	32943	38371	566	239	892	1189435	1158165	1220705	28689	26878	30500
26-07-2020	37737	34358	41115	579	249	910	1273972	1228364	1319580	30178	27802	32554
29-07-2020	39816	35664	43968	593	259	927	1358509	1296868	1420150	31666	28682	34649
01-08-2020	41895	36879	46912	607	270	945	1443046	1363842	1522250	33154	29523	36786
04-08-2020	43975	38016	49934	621	280	963	1527583	1429407	1625759	34643	30324	38961
07-08-2020	46054	39083	53026	635	290	980	1612120	1493661	1730579	36131	31088	41174
10-08-2020	48134	40087	56181	649	301	998	1696657	1556685	1836629	37619	31817	43422
13-08-2020	50213	41032	59395	663	311	1015	1781194	1618545	1943843	39108	32512	45704
16-08-2020	52293	41922	62664	677	322	1033	1865731	1679298	2052164	40596	33174	48018
19-08-2020	54372	42760	65985	691	332	1050	1950268	1738993	2161543	42084	33805	50364
22-08-2020	56452	43548	69355	705	343	1067	2034805	1797673	2271937	43573	34405	52741

In the above Table 1.6, we have seen and said that at the end of July more than 13.5 lakh peoples will be infected and daily cases reported approximately 40000 new cases. Whereas, due to this COVID approximately 600 people deaths per day and total deaths in India are more than 31000 recorded. In the third week of August 2020, total cases and total deaths will be more than 20 lakh and 40 thousand respectively due to this outbreak.

4 Discussions

The present chapter uses the Indian COVID-19 data from 14th March 2020 to 13th July 2020 for modeling purpose and obtains the forecast of the COVID 19 outbreak. We fitted three types of models using the historical data and identify the best model which can predict the future numbers of daily new cases, new deaths, total cases, and total deaths. The ARIMA time series model gives better forecasts as compared to other considered models. As per our long term forecast, this may help us to understand the actual situation in upcoming days and an advisory may be issued to people which should be followed for policy formation like “partially lockdown”, “Stay at home”, “home quartine”, “self-isolation” and “Work from home” to control the COVID-19 outbreak in India. Complimentary to a modeling approach to transmission dynamics of virus outbreaks, the data-driven based on dynamic as well as regression modeling approach provides real-time forecasting of daily new cases, new deaths, total cases, and total deaths for tracking, estimating the length of COVID-19 Outbreak. This can be an important tool to fight COVID 19 since we do not have an abundant testing facility in most of the states. So, in the future, more data and a healthier evaluation system can be made to handle this outbreak. However, the present chapter provides future information about the COVID-19 cases that can reach the same growth if the current situation cannot be changed. The study is also suggesting taking necessary steps for controlling the outbreak of COVID-19 in line with good plans.

5 Limitations

This article will invert-able make some limitations in terms of assumptions as per the construction of the model. When we find the order of the dynamic model as well as polynomial regression and the Exponential model for a certain period of COVID-19, we ignore the factor such as geographical conditions, population size, number of testing per day, recovered rate, Family income, etc. This paper is based on the recorded data for a specific period to fit and estimate the daily new cases, new deaths, total cases and total deaths of COVID-19; with the continuous release of outbreak data, these important indicators maybe play an important role in the spread of COVID-19 among the population.

References

1. Chatterjee, C., & Sarkar, R. R. (2009). Multi-step polynomial regression method to model and forecast malaria incidence. *PLoS One*, *4*(3), e4726.
2. Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 109864.
3. Jiang, X., Zhao, B., & Cao, J. (2020). Statistical analysis on COVID-19. *Biomedical Journal of Scientific & Technical Research*, 26(2), 19716-19727
4. Ma, L., & Zheng, J. (2017). A polynomial based model for cell fate prediction in human diseases. *BMC Systems Biology*, *11*(7), 126.
5. Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting*, *1*(2), 111-153.

6. Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modeling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, 101742.
7. Pandey, S., & Samantha, A. (2020). Real-time forecasting of COVID-19 prevalence in India using ARIMA model. International Journal of Management and Humanities (IJMH), 4(10), 78-83
8. Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020). COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. *Applied Sciences*, *10*(11), 3880.
9. Qi, H., Xiao, S., Shi, R., Ward, M. P., Chen, Y., Tu, W., Su, Q., Wang, W., Wang, X. & Zhang, Z. (2020). COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Science of the Total Environment*, 138778.
10. Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. *Chaos, Solitons & Fractals*, 109945.
11. Shastri, S., Sharma, A., Mansotra, V., Sharma, A., Bhadwal, A., & Kumari, M. (2018). A Study on Exponential Smoothing Method for Forecasting. *Int. J. Comput. Sci. Eng*, *6*(4), 482-485.
12. Song, X., Xiao, J., Deng, J., Kang, Q., Zhang, Y., & Xu, J. (2016). Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011. *Medicine*, *95*(26).
13. Sujath, R., Chatterjee, J. M., &Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, 34, 959–972.
14. Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International journal of Forecasting*, *19*(4), 715-725.
15. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, 138762.
16. Yadav, R. S. (2020). Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. *International Journal of Information Technology*, 1-10.
17. Yin, R., Luusua, E., Dabrowski, J., Zhang, Y., & Kwoh, C. K. (2020). Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. *Bioinformatics*, *36*(9), 2697-2704.

18. Yonar, H., Yonar, A., Tekindal, M. A., & Tekindal, M. (2020). Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. *EJMO*, *4*(2), 160-165.
19. Zhang, F., Kwoh, C. K., Wu, M., & Zheng, J. (2014, September). Data-driven prediction of cancer cell fates with a nonlinear model of signaling pathways. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 436-444.