Forecasting of COVID-19 in India Using ARIMA Model

Narayana Darapaneni
Director - AIML
Great Learning/Northwestern University
Illinois, USA
darapaneni@gmail.com

Deepak Reddy
Student – AIML
Great Learning
Bangalore, India
deepak8085@gmail.com

Anwesh Reddy Paduri
Research Assistant - AIML
Great Learning
Mumbai, India
anwesh@greatlearning.in

Pooja Acharya
Student – AIML
Great Learning
Bangalore, India
pooji3344@gmail.com

Nithin H S
Student – AIML
Great Learning
Bangalore, India
nitpappu@gmail.com

***Abstract*— The recent outbreak of COVID-19 in different states of India has major concerns for all administrative departments of the government and general public. The Pandemic has been tested positive in 1287945 individuals with 817209 recovered and 30601 succumbed to the disease. The first case of the novel coronavirus was detected in India on 30 January 2020. There was a lockdown imposed by the Government of India from 24 March 2020 and ended on 31 May 2020. A forecast in no lockdown scenario would help us to track the further progress of the disease and make sufficient data available in order to plan the future of hospital facilities, pharmaceutical investment etc.**

Keywords*— *COVID-19, Forecasting, lockdown, Time Series modelling, ARIMA

I. Introduction

COVID-19, an on-going epidemic, started in Wuhan city, China, in December 2019. The disease continues to cause infections in many countries around the world [1]. Considering the various aspects of transmission, the World Health Organization (WHO) declared it as a pandemic on 11 March 2020. Thereafter, COVID-19 has become a threat to human life and continues to de-stabilize various economic and social balance[13].

It has shown rapid infections in almost all countries, and there is no cure available for this deadly virus. The governments have issued precautionary measures such as social distancing, quarantine of suspected and infected cases, sanitization of streets and markets, and lockdown of the areas at different scales as a temporary measure.

In India, exponential growth has not been observed due to the implementation of lockdown by the central and state government. The data indicates that there is a strong relation to these measures, such as lockdown on the transmission behaviour of COVID-19.

On the other side, these measures have created substantial economic losses to the masses, and hence actions mentioned above cannot be imposed for a longer duration. Mainly, developing countries (such as India) cannot afford such payoff after some finite time.

The Indian government has continuously reviewed every state and has become more focused on localizing the lockdown. The need being in alarming states and few towns which are or possibly can be hotspots for COVID-19. For all these, it is important to have short-term forecasts which can be guiding light for decision-makers and administrations.

A data-based statistical model such as Autoregressive integrated moving average (ARIMA) [8][13] has proven effective in predicting short-term forecast including the dengue fever, haemorrhagic fever with renal syndrome and Tuberculosis in the past.

ARIMA has found to be more promising compared to similar models like the support vector machine and wavelet neural network. India being a diverse country, it will be essential to study the rate of infection of COVID-19 in different Indian states.

The current study aims to develop an auto-regressive integrated moving average (ARIMA) model to predict the COVID-19 patients' rise, recovery and death in India based on the daily data obtained from the Indian government [8] from 30 Jan 2020 to 3rd August 2020.

ARIMA models provide another approach to time series forecasting. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting and provide complementary approaches to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.

II. Materials and Methods

A. Data

For the validation and analysis purpose, the data was taken from COVID-19 in India [20]. The data comprises of daily counts of confirmed, cured and death cases in India till 3rd August 2020. Also, it comprises of additional information regarding state-wise testing for patients and the results segregated in positive and negative cases.

To understand the effects of lockdown on the rate of increase in the number of cases, the data used was after the first phase of lockdown, which had ended with few relaxations in economic and social activities. The count of the cases in the data before the lockdown would not be able to completely understand the impact of lockdown on the number of cases in the country.

B. ARIMA Model

During the analysis and forecasting of a time series, it is recommended to plot the time series data and analyse for

unique features. It may seem a tedious approach but provides proper insights on the spread and helps choose the modelling roadmap.[12]

Usually, while dealing with real-time data, most time series do not exhibit stationarity as they have no fixed mean. The ARMA model, introduced by Box and Jenkins, is the collection of popular methods that are directly applicable to modelling and analysing the time series [16]. The ARMA model is a combination of two models, the autoregressive AR (p) model and the moving average MA (q) model. These models are directly applicable to time series which are stationary. In the case of non-stationarity, differencing is applied to make it stationary. An ARMA model after differencing is known to be as ARIMA.

To achieve the above relevant values for p, d and q we used the grid search method using the auto_arima functionality provided by the python package, "pmdarima". Here it tries various sets of p and q parameters, selecting the model that minimizes the AIC (Akaike information criterion) to select the differencing terms, auto_arima uses a test of stationarity (such as an augmented Dickey-Fuller test) and seasonality (such as the Canova-Hansen test) for seasonal models. [18]

Figure 1 [16] gives a brief on how the method has been implemented. Where the orange region defines the automated grid search performed by auto_arima by lowering the AIC.

III. Results

The results are visualised in a phased manner of the lockdown viz.

1. Phase 1: 25th March to 14th April 2020
2. Phase 2: 15th April to 3rd May 2020
3. Phase 3: 4th May to 17th May 2020
4. Phase 4: 18th May to 31st May 2020

The data post 31st May 2020 is termed as "No Lockdown".

A. The goodness of fit

Table 1 represents the evaluation details of the model viz. R-squared, AIC, BIC (Bayesian information criteria) and MSE (Mean squared error) for the confirmed cases for India. The R-squared value of 0.994 suggests this be a good model with 0.06% unexplained variation. R-squared values range of 0.958 - 0.999 does imply a good fit.

TABLE 1: GOODNESS OF FIT

Model	India (Confirmed cases)
R-squared	0.994
AIC	1539.1
BIC	1551.6
MSE	749002.74

B. Confirmed cases

While going through the confirmed cases it was observed a seasonal component of 7 days which reduced the AIC by 100 points compared to a non-seasonal model. The forecast observed in the no lockdown phase indicates the benefits of lockdown. As of 3rd August 2020, 1874287 confirmed cases in total were predicted by the model against 1855332 actual cases. Table 2 details the forecasted result against the actual values for the daily count with the 95% confidence intervals.

TABLE 2: DAILY CONFIRMED CASES

Date	Forecast	95% Lower confidence	95% Upper confidence	Actual
25-Jul	47614	45858	49370	50072
26-Jul	51366	49010	53721	48932
27-Jul	54319	51812	56827	46484
28-Jul	52453	49625	55282	49637
29-Jul	53059	49715	56403	52479
30-Jul	59745	56024	63465	54968
31-Jul	63598	59549	67646	57486
01-Aug	63198	58228	68168	55117
02-Aug	67170	61406	72933	52672
03-Aug	70018	63729	76308	50789

Figure 2 gives details on the trend of daily counts of confirmed cases in India during various lockdown phases whereas figure 3 represents the cumulative sum of the same.

To establish a generic comparison among the top states in India with the maximum number of cases, cumulative plots for the states of Maharashtra, Andhra Pradesh, Tamil Nadu and Karnataka, are represented in Figure 4, Figure 5, Figure 6 and Figure 7 respectively. Each figure represents the lockdown phases to get a quick glimpse of the nature of trends in different timelines.

C. Deaths

The daily death counts do not show any significant rate of change in the lockdown period or without it. But an increase in mid of July indicates some rise which can have multiple aspects for the explanation. There was no seasonal component observed while analysing the model. But a significantly higher than normal deaths were reported on 17th June 2020 (2003 deaths) which was termed as an outlier and used without imputation.

Table 3 details of the forecasted result against the actual values and figure 8 briefly represents the data along with historical data for each lockdown phase.

TABLE 3: DAILY DEATHS FORECAST VS ACUTUAL

Date	Forecast	95% Lower confidence	95% Upper confidence	Actual
25-Jul	709	339	1079	703
26-Jul	709	332	1085	704
27-Jul	709	326	1092	642
28-Jul	709	320	1098	774
29-Jul	709	314	1104	775
30-Jul	709	308	1110	784
31-Jul	709	302	1116	764
01-Aug	709	296	1122	854
02-Aug	709	290	1127	760
03-Aug	709	285	1133	806

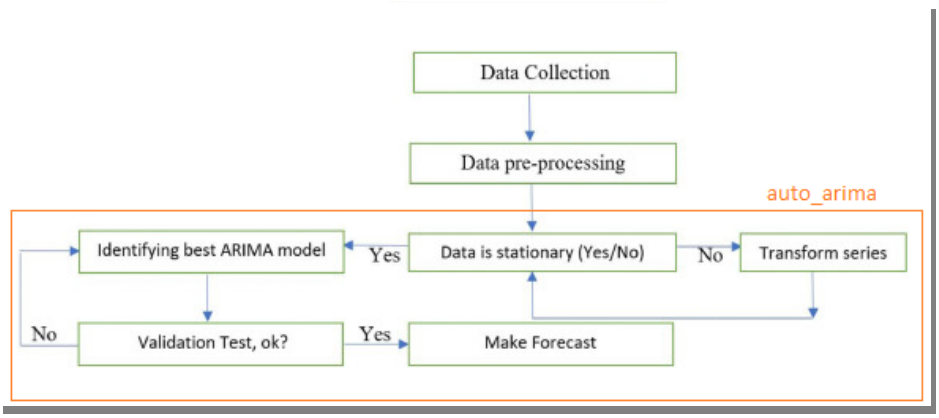

Figure 1: Flowchart for building an ARIMA model

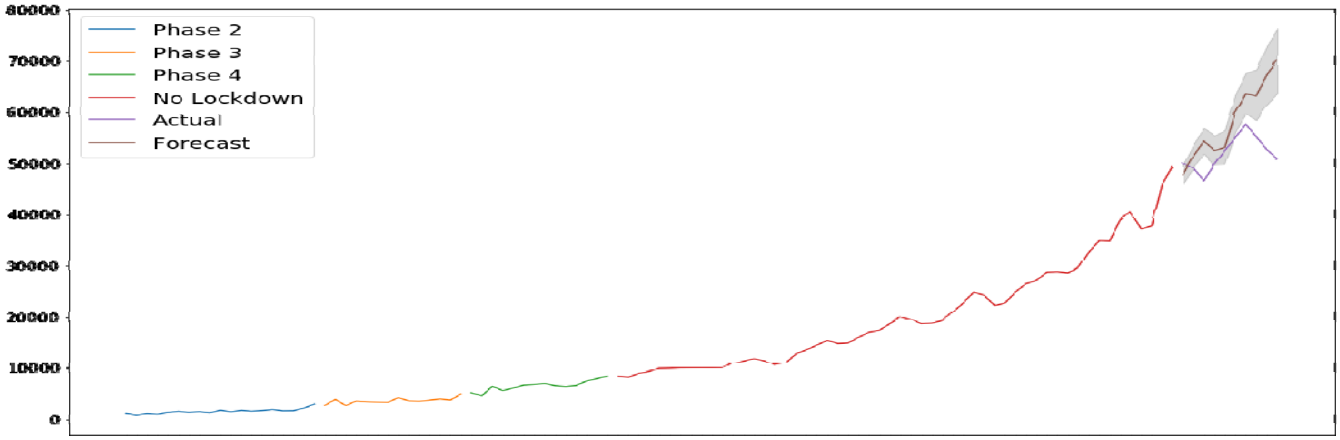

Figure 2: Daily confirmed cases, India

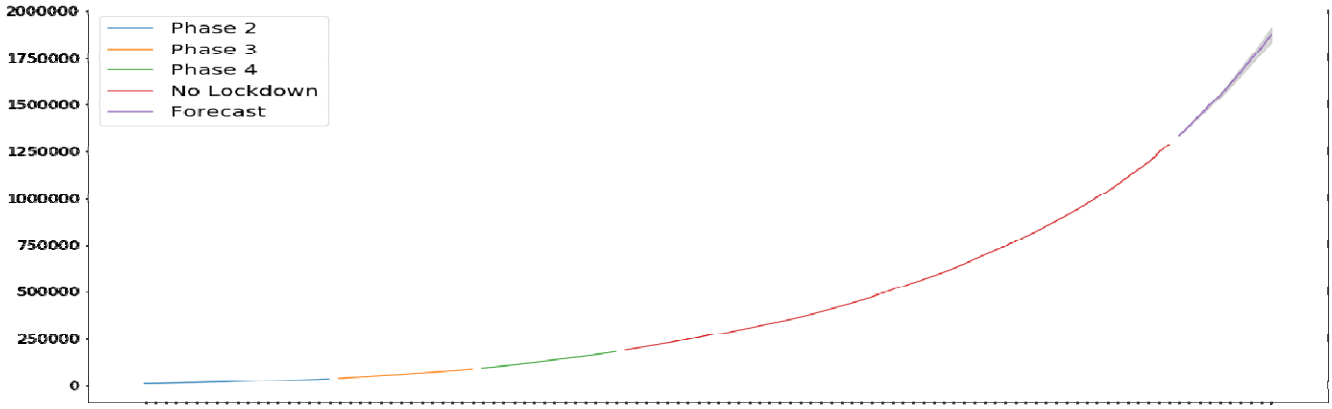

Figure 3: Cumulative confirmed cases, India

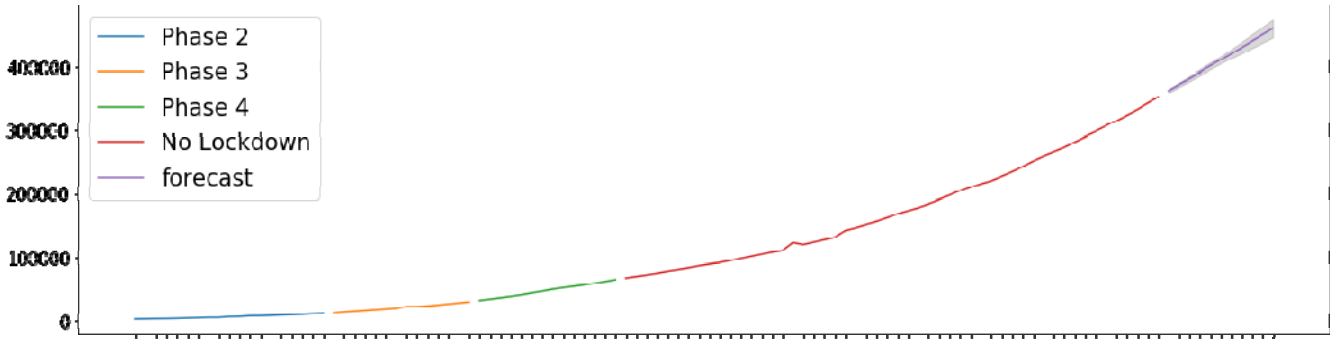

Figure 4: Cumulative confirmed cases, for the state of Maharashtra

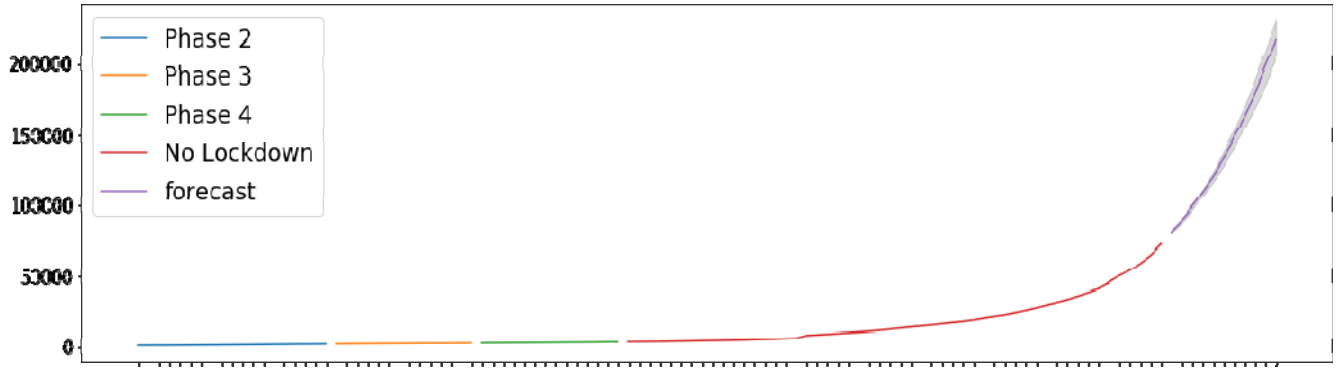

Figure 5: Cumulative confirmed cases, for the state of Andhra Pradesh

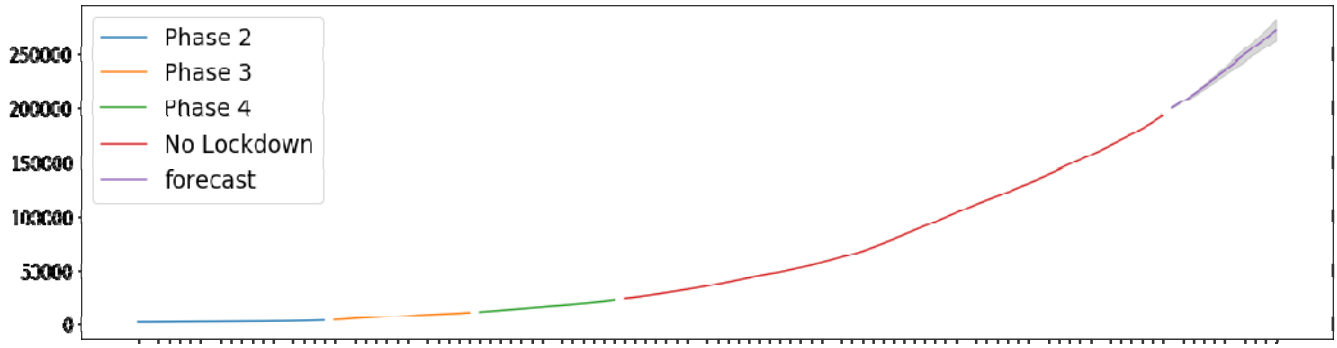

Figure 6: Cumulative confirmed cases, , for the state of Tamil Nadu

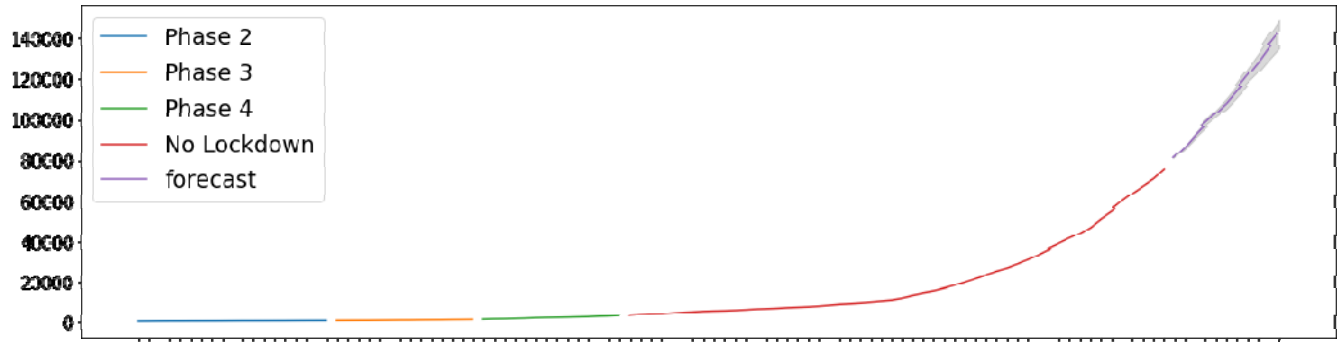

Figure 7: Cumulative confirmed cases, , for the state of Karnataka

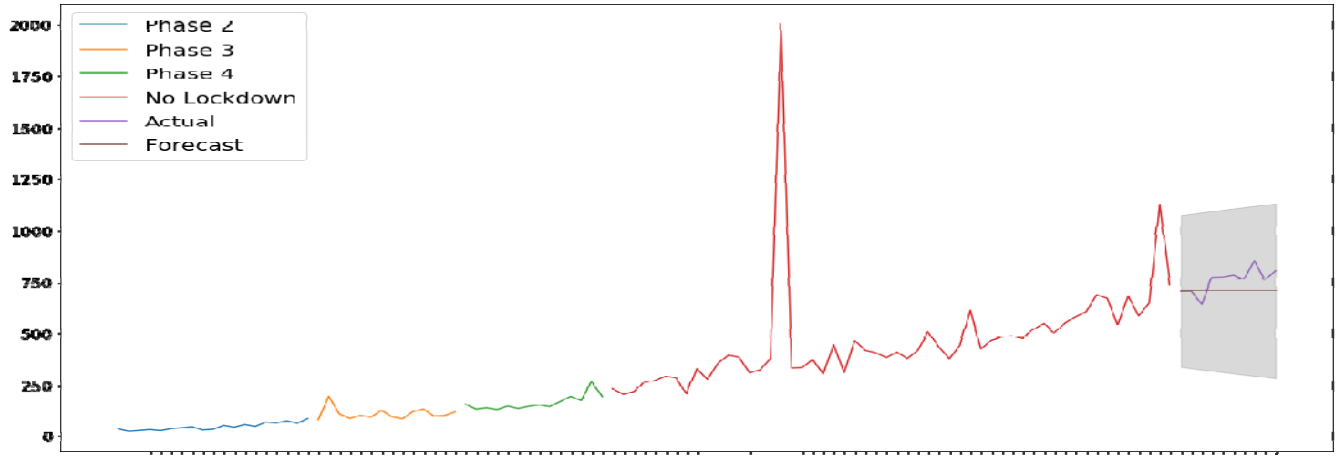

Figure 8: Daily Deaths, India

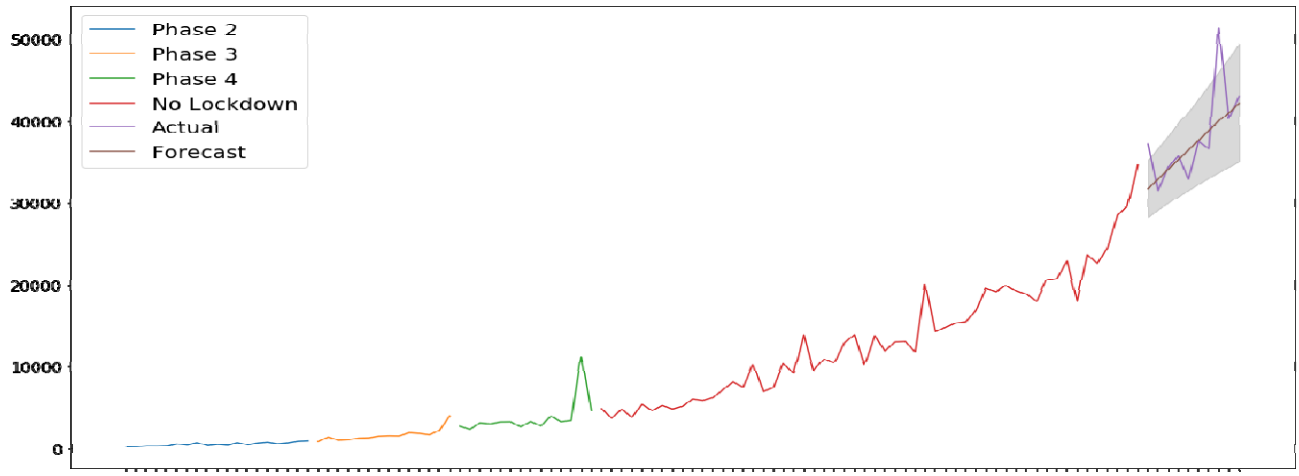

Figure 9: Daily Cured, India

D. *Cured*

The daily cured data had not shown any seasonal variations while analysing the model. It is highly analogous with the number of newly emerging cases as discussed earlier. There was no seasonal component observed while analysing the model.

Table 3 details of the forecasted result against the actual values and figure 9 represents the data with the phased lockdown data.

TABLE 4: DAILY CURED FORECAST VS ACTUAL

Date	Forecast	95% Lower confidence	95% Upper confidence	Actual
25-Jul	31732	28219	35245	37125
26-Jul	32898	29113	36683	31512
27-Jul	34065	29966	38163	34354
28-Jul	35232	30781	39682	35683
29-Jul	36398	31560	41236	32886
30-Jul	37565	32306	42824	37425
31-Jul	38731	33022	44441	36554
01-Aug	39898	33710	46086	51368
02-Aug	41065	34372	47757	40355
03-Aug	42231	35010	49452	43070

IV. DISCUSSIONS AND CONCLUSIONS

A. *Confirmed cases*

The rate of increase of confirmed cases was quite low during the lockdown phase. Lockdown eased the load on the medical infrastructure and helped prepare all regions for the future. Once the lockdown ended in India on 31 May 2020, the number of cases saw a dramatic increase approximately after 15-20 June. This being the tentative incubation time period of the virus in the human body starting 1st June 2020. The seasonal nature of the confirmed cases was observed during model building indicating a weekly pattern. Possible reason being that the medical teams were working on a lower capacity during one or two days of the week.

While observing figure 2 in detail, the actual test data was much lower than the forecast. One of the possible reason being the higher AIC value incurred during the modelling. Other reasons might be possible benefits of government actions taken earlier to reduce the infection, public behaviour, tracing and time quarantine activities etc.

B. *Comparison of confirmed cases among States*

While comparing the cumulative graph of cases in the top four states in India, unique variations were observed. In Maharashtra & Tamil Nadu the cases were on the rises since the end of Phase 2 of lockdown. The reasons may be due to increased movement of people with lenient lockdown protocols to reduce the economic losses incurred in the past. Also, they comprise of the huge migrant workforce from various parts of the country which completely rely on the economic activities of these states.

The states of Andhra Pradesh and Karnataka saw the cases on the rise after the end of lockdown and might have taken the authorities by surprise due to the sudden jump in cases. This kind of sudden rise might be the unchecked growth of confirmed cases and lack of containment activities.

C. *Deaths*

There we possibilities of the deaths being under-reported as we could observe a single-day spike in the data, which may be accumulated deaths from the past days. This indicates towards some factors like the workload of the medical authorities, prioritizing the patient's health before documentation, medical staff being infected and more staff involved in contact tracing and containment activities. Deaths occurring in India looks to have a mild increase even in the post lockdown period showing better resistance of the masses against the disease. The prediction to supports the hypothesis.

D. *Comparison with the modified logistic growth model.*

On comparison with a related model [23], the Modified logistic growth model, where the analysis was done on data before the lockdown was lifted, we found a better AIC and R-squared values (Table 5). The common reason for the observation could be due to a slower growth rate of the infection in the country during the lockdown showing a substantial improvement in the forecast.

TABLE 5: GOODNESS OF FIT (MODIFIED LOGISTIC GROWTH MODEL

Model	India (Confirmed cases)
R-squared	0.997
AIC	663.6
BIC	1337.1
MSE	575419.8

E. Limitations of ARIMA

Results clearly indicate that the predictions are well within the range of 95% confidence interval, except of the confirmed cases. But the main advantage of ARIMA forecasting approach is surely its ease of application and interpretation. By the contrary, it is sensitive to outliers in the data and, do not account for the noise, that is unknown by definition. For these reasons, it may be considered a good model for short-term forecasting, but the results should be interpreted with prudence.

REFERENCES

[1] T. Sathish, A. Ray, and N. Nanda Gopal, "Predictions of COVID-19 patientsraise, recovery and death rate in India by ARIMA model."

[2] V. Jha, "Forecasting the transmission of Covid-19 in India using a data driven SEIRD model," Arxiv.org. [Online]. Available: http://arxiv.org/abs/2006.04464v1.

[3] A. Kakar and S. Nundy, "COVID-19 in India," J. R. Soc. Med., vol. 113, no. 6, pp. 232–233, 2020.

[4] B. Banerjee, P. K. Pandey, and B. Adhikari, "A model for the spread of an epidemic from local to global: A case study of COVID-19 in India," arXiv [physics.soc-ph], 2020.

[5] K. Chatterjee, K. Chatterjee, A. Kumar, and S. Shankar, "Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model," Med J. Armed Forces India, vol. 76, no. 2, pp. 147–155, 2020.

[6] D. K. Rajendran, V. Rajagopal, S. Alagumanian, T. Santhosh Kumar, S. P. Sathiya Prabhakaran, and D. Kasilingam, "Systematic literature review on novel corona virus SARS-CoV-2: a threat to human era," Virusdisease, vol. 31, no. 2, pp. 161–173, 2020.

[7] A. K. Rauta, Y. S. Rao, and J. Behera, "Spread of COVID-19 in Odisha (India) due to Influx of Migrants and Stability Analysis using Mathematical Modelling," Researchsquare.com. [Online]. Available: https://www.researchsquare.com/article/rs-34007/latest.pdf.

[8] S. P. Marbaniang, "Forecasting the Prevalence of COVID-19 in Maharashtra, Delhi, Kerala, and India using an ARIMA model."

[9] R. Muthusami and K. Saritha, "Statistical analysis and visualization of the potential cases of pandemic coronavirus," Virusdisease, vol. 31, no. 2, pp. 204–208, 2020.

[10] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," Data Brief, vol. 29, p. 105340, 2020.

[11] "Coronavirus in India: Latest map and case count," Covid19india.org. [Online]. Available: https://www.covid19india.org/. [Accessed: 25-Jul-2020].

[12] N. Poonia and S. Azad, "Short-term forecasts of COVID-19 spread across Indian states until 1 May 2020," arXiv [q-bio.PE], 2020.

[13] "Chapter 8 ARIMA models | forecasting: Principles and practice," Otexts.com. [Online]. Available: https://otexts.com/fpp2/arima.html.

[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and control. John Wiley & Sons, 2015.

[15] S. E. Said and D. A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order," Biometrika, vol. 71, no. 3, pp. 599–607, 1984.

[16] F. M. Khan and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India," Journal of Safety Science and Resilience, vol. 1, no. 1, pp. 12–18, 2020.

[17] S. Prabhakaran, "ARIMA model - complete guide to time series forecasting in python | ML+," Machinelearningplus.com, 18-Feb-2019. [Online]. Available: http://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python.

[18] "pmdarima.arima.auto_arima — pmdarima 1.7.1 documentation," Alkaline-ml.com. [Online]. Available: https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html

[19] J. M. Portilla, "Using Python and auto ARIMA to forecast seasonal time series," Medium, 26-Mar-2018. [Online]. Available: https://medium.com/@josemarcialportilla/using-python-and-auto-arima-to-forecast-seasonal-time-series-90877adff03c. [Accessed: 10-Sep-2020].

[20] SRK, "COVID-19 in India." .

[21] Sangarshanan, "Time series Forecasting — ARIMA models - Towards Data Science," Towards Data Science, 03-Oct-2018. [Online]. Available: https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06.

[22] Wikipedia contributors, "COVID-19 pandemic lockdown in India," Wikipedia, The Free Encyclopedia, 10-Sep-2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_lockdown_in_India&oldid=977661921.

[23] B. Malavika, S. Marimuthu, M. Joy, A. Nadaraj, E. S. Asirvatham, and L.Jeyaseelan,"Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models," Clin. Epidemiol.Glob.Health,2020