Demography India
Vol. 49, Special Issue (2020), pp. 66-75 ISSN 0970-454X

Modelling and Forecasting of COVID-19 cases in Odisha and India

Prafulla Kumar Swain[*1], Manas Ranjan Tripathy[2], Diptismita Jena[3], Haile Mekonnen Fenta[4] and Dereje Tesfaye Zike[5]

Abstract: The novel coronavirus disease (COVID-19) has shaken the entire world in its devastating annihilation of humanity. The alarming increase in number of confirmed COVID cases in India requires an urgent step to monitor and control this spread. Thus, it is inevitable to develop a model which can predict future confirmed and death cases. Time series models are significant in predicting the impact of the COVID-19 outbreak. In this paper we have developed an Auto Regressive Integrated Moving Average (ARIMA) model to predict the number of COVID 19 cases for India and Odisha. The model prediction suggests that an increasing trend would be continuing for both Odisha and India as whole for next two weeks. The forecasted values are in good agreement with actual cases in both the scenario. These findings would be useful for government in formulating policy related to health care systems so that the system and medical professions can better prepared to combat the pandemic.

Keywords: Coronavirus, COVID-19, ARMA, ARIMA, Forecasting.

Introduction

The World Health Organization (WHO) has declared the COVID-19 as a global pandemic on 11th March 2020 as it has been spreading all over the world in an alarming rate with severity. The COVID-19 outbreak is highly similar to the Severe Acute Respiratory Syndrome (SARS) outbreak occurred in 2003. The transmission potential of COVID-19 and its epidemiological features are still unclear (Mizumoto, et al., 2020). As of 31st May, 2020 there are 6,057,853 confirmed cases and 371,166 deaths have been reported worldwide (WHO report, 2020). The pandemic has a profound impact on people's life, economy and health care system. It has also created an unprecedented challenge for government to tackle this crisis.

In India the very first COVID-19 case was identified towards the end of January 2020 in Kerala, when 3 students returned from Wuhan, China. So far India's response to the outbreak is noteworthy. As of 31stMay 2020, India has more than 1, 82,143 cases with more than 5,607 deaths (MoHFW report). India has taken various steps (including complete lockdown of the entire country) to prevent the spread of new infections. In order to have a suitable plan for

* Corresponding Author
[1] Assistant Professor, Department of Statistics, Utkal University, Bhubaneswar, India, Email: prafulla86@gmail.com
[2] Research Scholar, Department of Statistics, Ravenshaw University, Cuttack, India, Email: manasranjantripathy402@gmail.com
[3] Research scholar, Department of Statistics, Ravenshaw University, Cuttack, India, Email: diptismita.jena@gmail.com
[4] Assistant Professor, Department of Statistics, College of Science, Bahir Dar University, Ethiopia, Email: hailemekonnen@gmail.com
[5] Assistant Professor, Department of Statistics, College of Science, Bahir Dar University, Ethiopia, Email: Email id: derejetesfaye11@gmail.com

COVID-19, especially for controlling the outbreak, forecasting the future confirmed cases are critical. These forecast values can assist in decision making and logistical planning in healthcare systems. Some researchers have used statistical and mathematical modelling. Authors have also projected that India may vary between 97 thousand to 250 million infected cases and 30 thousand deaths due to COVID-19 by July 2020 using prediction models (CDDEP, 2020).

It is important to develop the state and region-specific model for COVID-19 cases so that effective strategies can be drawn for the region. Considering the eastern Indian state of Odisha, the first confirmed case (a returnee from Italy) was emerged on 16th March 2020. The state has able to managed the COVID-19 outbreak well till first week of May. However, the cases started increasing gradually once the migrants returned to the other states. As of 31stMay 2020, there are 2,104 confirmed cases and only 9 deaths so far. The state has the distinction of low rate of mortality and high recovery rate. Being a poor state and grossly lacking in health infrastructure has managed the pandemic very effectively. The experience of Odisha in disaster management provides an advantage to respond the COVID-19 pandemic. The state government has also declared the pandemic as state disaster under the provision of disaster management act 2005. Since the mechanisms of COVID-19 spreading are not completely understood, the number of infected people is large and the effects of containment are evaluated essentially on an empirical basis. And also, sudden increase in number of cases may collapse the health care system. Thus, it is imperative to study the trend and forecast the confirmed and death cases in daily basis.

Recently different Statistical models have been used to predict COVID-19 cases, viz., Lin et al.,2020 used a SEIR (Susceptible- Exposed-Infectious- Removed) model for the spread in China. Anastassopoulou *et.al.,* 2020 used a SIRD (Susceptible- Infected- Recovered-Dead) model to estimate the basic reproduction number (R0), and the per day infection mortality and recovery rates. Silva *et. al.,* 2020 used a growth curve model for total cases in Brazil using a Bayesian approach.

Several studies in the literature support the use of the time series modelling, particularly ARIMA model (Ceylan, 2020), TP–SMN–AR model (Maleki *et.al.*, 2020), Holt-Winters models (Sharma and Nigam, 2020), Holt's Trend (Gupta and Pal, 2020). Curve Estimation and Exponential smoothing model (Yonar *et.al*., 2020).

At present, there is neither a treatment nor a vaccination for the COVID-19 infection. Modelling and prediction of prevalence of infected cases and epidemiological characteristics are important issues in planning and formulating health care systems and for effective communication with the public. Sometimes the expected numbers are useful in taking immediate and effective steps in right directions. In this work we have made an attempt to forecast the COVID-19 infected cases for India along with the state of Odisha using Automatic Regressive Integrated Moving Average (ARIMA) model. The ARIMA model has been successfully applied in the field of health as well as in different fields in the past due to its simple structure, fast applicability and ability to explain the data set (Cao *et. al.,* 2020).

Materials and Methods

Data Sources

The time series data on cumulative confirmed cases, death and recovery cases for India have been extracted from WHO situation report and University of Oxford website https://ourworldindata.org/coronavirus and for Odisha, data have been extracted from www.health.odisha.gov.in.site during the period of 30th Jan, 2020 to 31st May 2020. Since these data are open source and publicly available so there is no need for ethical clearance.

Methods

The ARIMA model was proposed by George Box and Gwilym Jenkins in the year 1970s (Box *et. al.*, 2015). The ARIMA is one of the most used time series models as it takes into account changing trends, periodic changes and random disturbances in the time series. ARIMA is suitable for all kinds of data, including trend, seasonality and cyclicity. ARIMA modelling consists of four vital steps viz. identification of potential models, estimation of parameters in that potential models, diagnostic checking of residuals for white noise and forecasting by taking the help of selected model.

ARIMA model is generally referred to as an ARIMA (p,d,q) where p signifies the order of auto regression, d denotes the degree of difference, and q is the order of moving average. In other words, the p and q were the number of significant lags of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots, respectively, and d was the different order needed to remove the ordinary non-stationarity in the mean of the error terms. ARIMA is a joint model of two models, autoregressive AR(p) and moving average MA(q) and is integrated using the difference variabled. The general formula of AR (p) and MA (q) models can be expressed in Eqs. (1) and (2), respectively.

An autoregressive AR(p) model of order p can be written as:

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \ldots\ldots\ldots\ldots + \emptyset_p y_{t-p} + e_t \qquad (1)$$

Where c is a constant, e_t is a white noise ($e_t \sim N(0, \sigma^2)$, $\emptyset = (\emptyset_1, \emptyset_2, \ldots\ldots, \emptyset_p)$; is the vector of model coefficients & p is a non-negative integer.

A moving average MA(q) model of order q uses past forecast errors in a regression model as

$$y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots \ldots\ldots\ldots - \theta_q e_{t-q} \qquad (2)$$

Where c is a constant, e_t is a white noise ($e_t \sim N(0, \sigma^2)$, $\theta = (\theta_1, \theta_2, \ldots\ldots, \theta_p)$is the vector of model coefficients & q is a non-negative integer.

The ARMA (p, q) process of orders p and q is defined as

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \ldots + \emptyset_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$
$$\Rightarrow y_t = c + \sum_{j=1}^{p} \emptyset_i y_{t-i} - \sum_{j=1}^{q} \theta_j e_{t-j} + e_t \qquad (3)$$

ARIMA (p, d, q) model can be written as: $y_t - 2y_{t-1} - \cdots - y_{t-d} = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \ldots + \emptyset_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$ (4)

Where, p autoregressive terms, d is the non-seasonal differences, q is the number of lagged forecast errors.

The best model is selected based on the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values. The smaller the size, the better was the model. Generally, the AIC is calculated using the relation $AIC = 2k - 2\log(L)$ and $BIC = k * log(n) - 2\log(L)$; Where k(=p+q+1) is the number of parameters in the statistical model and L is the maximized value of the likelihood function for the estimated model. All the analyses and the forecast were computed using the R statistical software package. The statistical significance was decided at $p < 0.05$.We have applied ARIMA model to the time series data of confirmed COVID-19 cases in Odisha and India.

Results and Discussions

As discussed in the methodology, we have applied ARIMA model to Odisha and India dataset for forecasting. It can be clearly observed from the Figure 1(top) that, there is an increasing growth of COVID-19 patients in India which is non-stationary in nature. So, to make it stationary by stabilizing the mean, we need to take its first difference. Again the first difference in Figure 1 (bottom left) is non-stationary in nature as its increasing trend is not eliminated and its Dickey-Fuller test (t=1.5775, p=0.99) is not significant. Now taking second differencing as shown in fig.1 (bottom right) shows a stationary series and its Dickey-Fuller test (t=-6.787, p=0.01) is significant at 5% level of significance. Thus, we have identified the degree (d) of ARIMA model is 2. Then from ACF and PACF plots for COVID-19 data for India (Figure 3), the following competing models have been identified for the estimation of parameters : ARIMA(1,2,2), ARIMA(0,2,2), ARIMA(1,2,1), ARIMA(0,2,1), ARIMA(0,2,0). The best fit model can be identified by comparing AIC and BIC values. The ARIMA (p,d,q) model with lowest Akaike Information Criterion (AIC) and Bayesian information criterion (BIC), can be treated as best model. The ARIMA (0,2,2) model taken as best fit model for India COVID-19 data as its AIC (AIC=1697.44) and BIC (BIC=1705.80) are lowest among all the suggested models (Table 1).

Table 2 shows the value of estimated parameters with its significance values. All the estimated values are significant at 95% CI and considerably different from zero. The forecasted values with its 95% forecasted intervals are reflected in Table3. As per forecasted values for the next 15 days the COVID-19 confirmed cases are expected to lie between 287857 and 334505 for India. The forecasted trend with its 95% prediction interval for India is shown in Figure 5. Thus, it is extremely crucial that people are made aware of the situation and the lockdown and social distancing should strictly imposed in the whole country to prevent further transmission of the infection.

Figure 1: Trend of COVID-19 patients in India (top), first order difference India (bottom left) and second order difference of COVID-19 patients in India (bottom right)

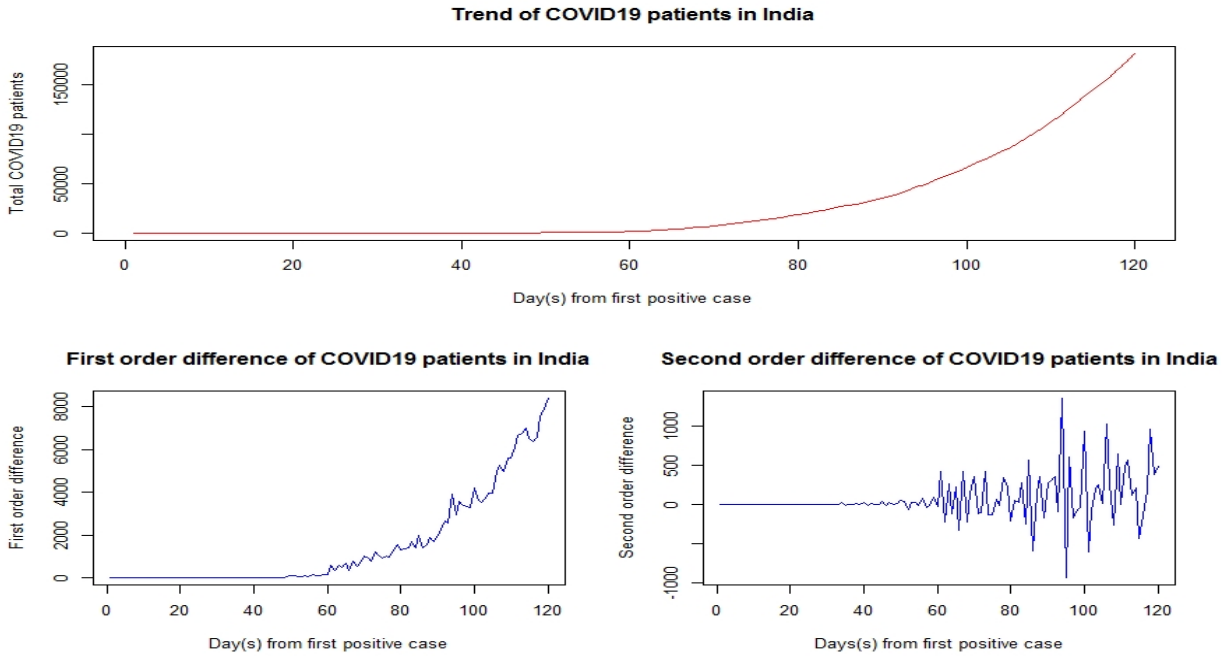


Figure 2: Trend of COVID-19 patients in Odisha (top), first order difference (bottom left) and second order difference of COVID-19 patients in Odisha (bottom right)

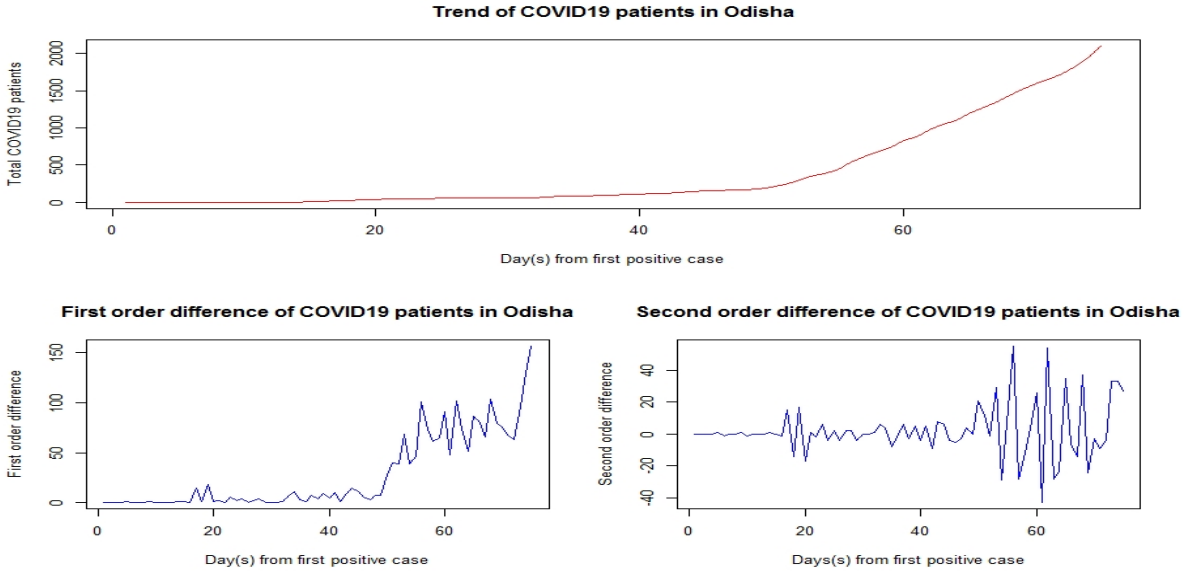


Figure 3: The ACF and PACF plots for IndiaCOVID-19 data

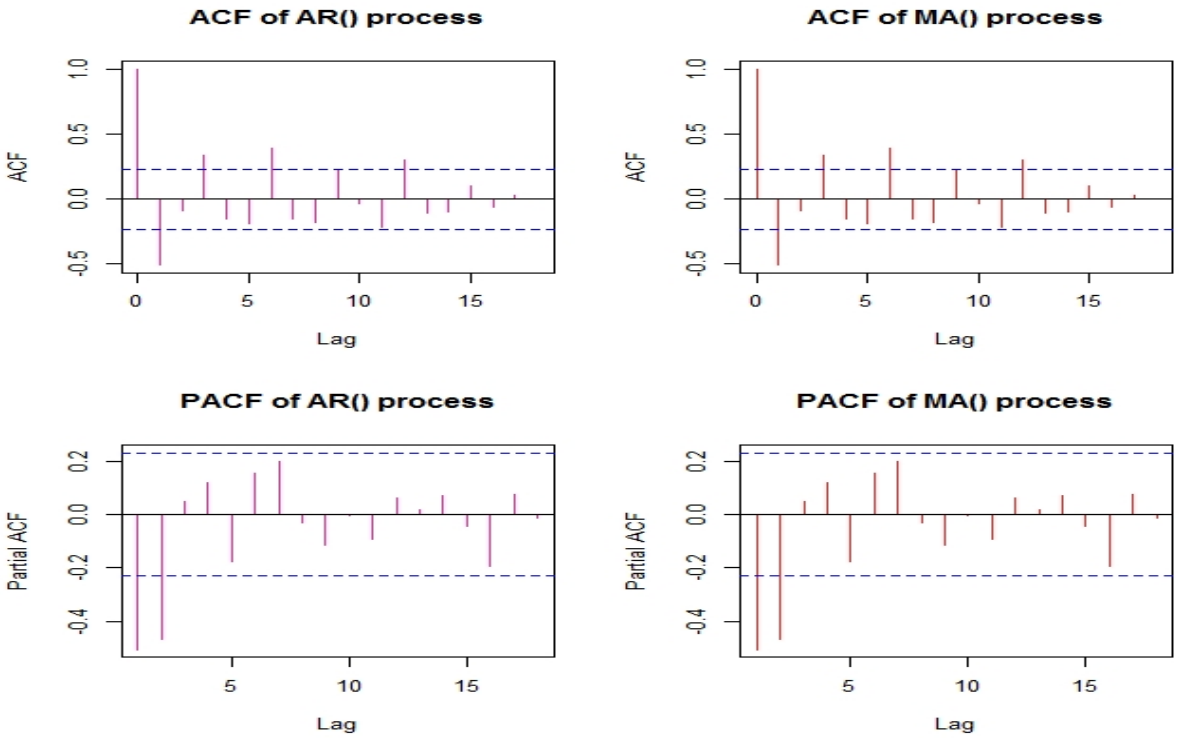


Figure 4: The ACF and PACF plots for OdishaCOVID-19 data

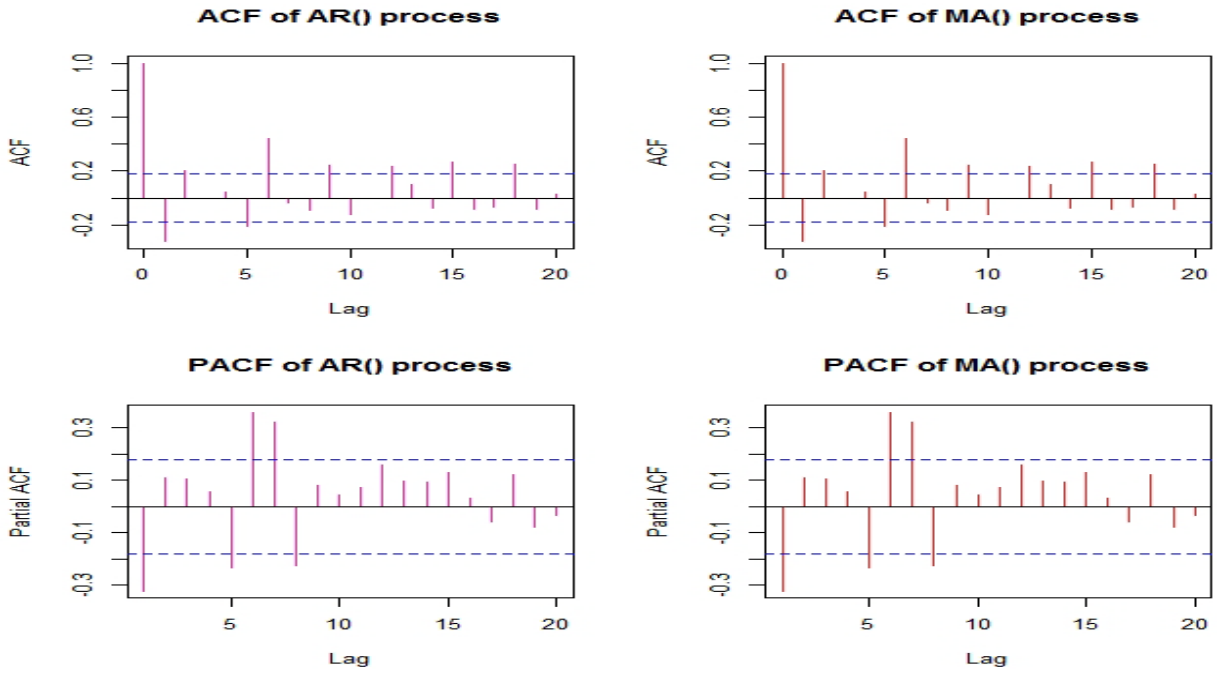


Table 1: AIC for ARIMA models for Covid-19 cases, Odisha and India

	Model	Loglikelihood	AIC	BIC
India	ARIMA (1,2,2)	-845.70	1699.39	1710.54
	ARIMA (0,2,2)	**-845.72**	**1697.44**	**1705.80**
	ARIMA (1,2,1)	-849.53	1705.05	1713.42
	ARIMA (0,2,1)	-852.08	1708.15	1713.73
	ARIMA (0,2,0)	-854.82	1711.64	1714.42
Odisha	ARIMA (2,2,1)	-310.35	628.70	637.97
	ARIMA (2,2,0)	-310.87	627.75	634.70
	ARIMA (1,2,1)	-311.85	629.70	636.65
	ARIMA (1,2,0)	-312.96	629.92	634.56
	ARIMA (2,2,2)	**-306.52**	**623.05**	**634.36**

Table 2: Parameters of ARIMA models

	Parameters	Coefficients	Std. error	t-statistic	P value
India	MA1	-0.2170	0.0710	-3.06	0.020
	MA2	0.4987	0.1519	3.28	0.000
Odisha	AR1	-0.7924	0.1157	-6.85	0.000
	AR2	-0.8384	0.0961	-8.72	0.001
	MA1	0.4822	0.1469	3.28	0.000
	MA2	0.7159	0.1290	5.55	0.001

Table 3: Forecasting of total confirmed cases of COVID-19 for the next fifteen days based on ARIMA models with 95% confidence interval

Date	India				Odisha			
	Actual	Forecast	Lower Limit	Upper Limit	Actual	Forecast	Lower Limit	Upper Limit
01-06-2020	190535	190746	190198	191295	2245	2255	2226	2284
02-06-2020	198706	199349	198227	200471	2388	2410	2352	2466
03-06-2020	207615	207951	205930	209973	2478	2566	2474	2657
04-06-2020	216919	216554	213427	219680	2608	2717	2581	2853
05-06-2020	226770	225156	220762	229551	2781	2871	2689	3054
06-06-2020	236657	233759	227955	239563	2856	3027	2794	3260
07-06-2020	246628	242361	235020	249702	2994	3179	2890	3469
08-06-2020	256611	250964	241969	259958	3140	3333	2984	3682
09-06-2020	266598	259566	248808	270324	3250	3489	3078	3900
10-06-2020	276583	268169	255545	280792	3386	3642	3164	4120
11-06-2020	286579	276771	262185	291357	3498	3795	3247	4343
12-06-2020	297535	285374	268732	302015	3723	3950	3331	4570
13-06-2020	308993	293976	275191	312761	3909	4104	3408	4799
14-06-2020	320922	302579	281565	323592	4055	4257	3483	5031
15-06-2020	332424	311181	287857	334505	4163	4412	3557	5267

Figure 5: The forecasted graph for COVID-19 patients in India

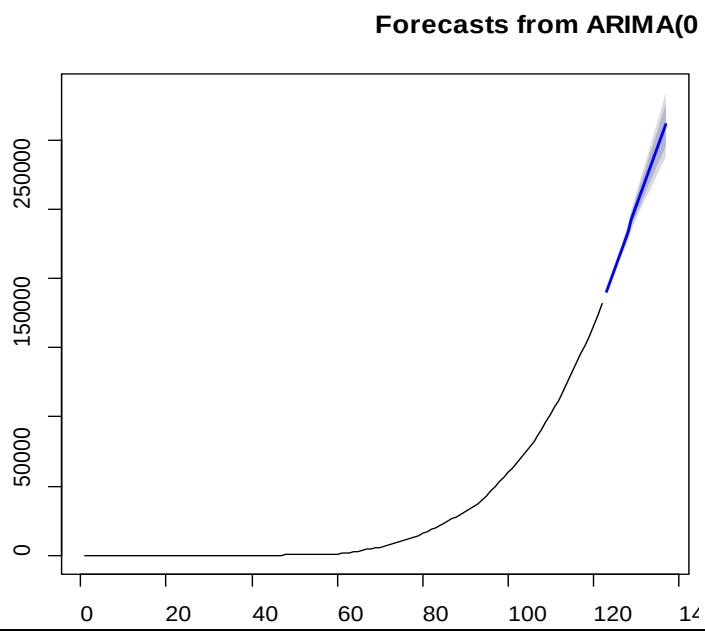
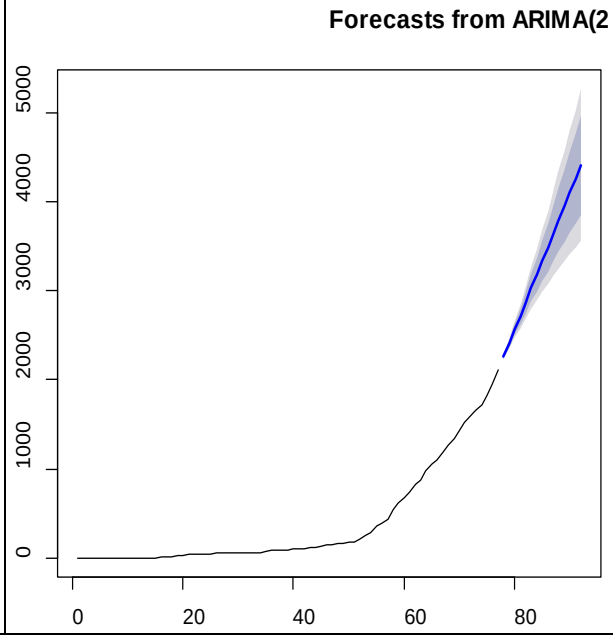


Figure 6: The forecasted graph for COVID-19 patients in Odisha



On the other hand, Figure 2 (top) shows a sharp increasing growth of COVID-19 patients in Odisha which is non-stationary in nature. To make it stationary we need to take its first difference which will stabilize its mean. After taking the first order difference as shown in Figure 2 (bottom left), again we have it is non-stationary in nature as its increasing trend still present and its Dickey-Fuller test (t=-1.3195, p=0.8533) doesn't show any significant result. Now second order differencing as shown in Figure 2 (bottom right) gives some extent to stationarity in the series and its Dickey-Fuller test (t=-4.4297, p=0.01) is significant. In this way we have identified the degree (d) for the ARIMA model is 2. Then from ACF and PACF plots for COVID-19 data for Odisha (Figure 4), the following competing models have been identified for the estimation of parameters: ARIMA(2,2,1), ARIMA(2,2,0), ARIMA(1,2,1), ARIMA(1,2,0), ARIMA (2,2,2). The best fit model can be identified by comparing AIC and BIC values. The ARIMA (2,2,2) model taken as best fit model for Odisha COVID-19 data as its AIC=623.05and BIC=634.36 are lowest among all the suggested models (Table1). The forecasted values with its 95% forecasted intervals are reflected in Table 3. As per forecasted values for the next 15 days the COVID-19 confirmed cases expected to lie between 3557 and 5267 for Odisha. Figure 6 shows the forecasted trend with its 95% prediction interval for Odisha. By looking at the forecasted graph we can expect increasing trend in COVID-19 cases for the next couple of weeks. Thus, government of Odisha should focus on sensitization of people about the seriousness of COVID 19 and also ensure the quarantine of all the people coming from outside the state. This is an alarming state of affairs for the policymakers to practice better resources in prevention and control of the further transmission. Our findings may help them to plan the required strategy for supplying potential health facilities to all individuals.

We have also compared the forecasted values with actual confirmed cases till June 15, 2020 to assess the potential of our model. It shows a good agreement; particularly the forecasted values of Odisha were so close to actual confirmed COVID-19 cases. To estimate model adequacy, we have also performed residual analysis for both the selected model and it was observed that the residuals are Independent and Identically Distributed (I.I.D.) and uncorrelated (ACF and PACF plot are not shown here). It is important to note here that accurate prediction of COVID-19 cases is really challenging as a number of hidden factors are involved viz. individual behaviour, human tendency, virology and human immune system etc. So, effective strategies are needed to contain the spread of the disease. We have not taken into consideration several other potential factors viz., medical facilities, climatic conditions, socio-economic status etc., in our prediction that may affect the spreading ofCOVID-19.

Conclusion

In this paper we have forecasted the COVID-19 patients for 15 days in India and Odisha. From the outcomes of this study it is expected that the COVID-19 positive cases will increase for the upcoming days in India and Odisha as well. This result will be helpful and support for the government to make planning and strategy to handle the rapidly growing pandemic for the next couple of weeks. However, the lockdown, city sanitization, isolation and quarantine may affect this forecast. For future prospective we are also planning to extend this work and use different multivariate models for forecasting, which will provide a further precise description of the current pandemic situation.

Acknowledgements

The authors are grateful to the editor and reviewers for their constructive suggestions.

References:

Anastassopoulou C. et.al, 2020, Data-based analysis, modeling and forecasting of the COVID-19 outbreak. *PLoS ONE,* 15(3): e0230405.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015, Time Series Analysis: Forecasting and Control, 5th Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.

Cao, L. Ting, Liu, H. Hui, Li, J., Yin, X. Dong, Duan, Y., Wang, J., 2020, Relationship of meteorological factors and human brucellosis in Hebei province, China. *Sci. Total Environ,*703, 135491. https://doi.org/10.1016/j.scitotenv.2019.135491.

CDDEP, 2020, COVID-19 Modeling with India SIM. India State-Level Estimates, Available from: https://cddep.org/covid-19/

Ceylan, Z., 2020, Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of the Total Environment 729 (2020): 138817.

Gupta, R. and Pal, S.K., 2020, Trend Analysis and Forecasting of COVID-19 outbreak in India.medRxiv preprint doi: https://doi.org/10.1101/2020.03.26.20044511

Maleki, M. et.al, 2020, Time series modelling to forecast the confirmed and recovered cases of COVID-19.Travel Medicine and Infectious Disease, https://doi.org/10.1016/j.tmaid.2020.101742

Mizumoto K, Kagaya K and Chowell G., 2020, Early epidemiological assessment of the transmission potential and virulence of 2019 Novel Coronavirus in Wuhan City: China, 2019–2020. medRxiv 2020.

MoHFW, https://www.mohfw.gov.in

Petropoulos F, Makridakis S., 2020, Forecasting the novel coronavirus COVID-19. *PLoS ONE*, 15(3): e0231236. https://doi.org/10.1371/ journal.pone.0231236.

Sharma, V.K. and Nigam, U., 2020, Modelling of Covid-19 cases in India using Regression and Time Series model,medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540

Silva, R.R. et.al, 2020, A Bayesian analysis of the total number of cases of the COVID-19 when only a few data is available, A case study in the state of Goias, Brazil, medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20071852

WHO, 2020, Coronavirus disease 2019 (COVID-19) Situation Report –130, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200529-covid-19-sitrep-130.pdf?sfvrsn=bf7e7f0c_4

Yonar, H. et.al, 2020, Modelling 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.