J. Inst. Eng. India Ser. B (December 2021) 102(6):1303–1309
https://doi.org/10.1007/s40031-021-00546-0

CASE STUDY

Predicting the New Cases of Coronavirus [COVID-19] in India by Using Time Series Analysis as Machine Learning Model in Python

Vikas Kulshreshtha[1] · N. K. Garg[2]

Received: 10 August 2020 / Accepted: 18 January 2021 / Published online: 18 February 2021

Abstract Today world is going through a critical phase. The whole world is infected from the coronavirus [COVID 19]. In India also the number of new cases keeps on increasing. In this paper, the machine learning model has been developed using time series analysis (ARIMA model) for predicting the new cases in India in the next coming days. In this work, results are also compared with the predictive values generated from the ARIMA and AR model and concluded that the ARIMA model is the best fit model as compared to AR model for predicting the new cases in India. Python programming language has been used for implementation. The dataset from January 1, 2020 to July 31, 2020 has been taken for analysis. This paper is useful for researchers for further analysis of COVID-19 pandemic in India.

Keywords COVID-19 · Cases · Pneumonia · AR model · ARIMA model · Pandemic · Python

Introduction

The coronavirus case was emerged in Wuhan city, China in December 2019 [1]. Initially, anonymous pneumonia case was reported which was studied by respiratory samples and experts declared as pneumonia. It was later known as coronavirus pneumonia due to a novel coronavirus [2]. The World Health Organization officially stated the disease 'COVID-19.′ It spreads very fast globally. There were 17,106,007 confirmed cases and 668,910 deaths globally as of July 31, 2020 [3]. COVID-19 becomes global threat for public health [4]. In India, the first active case was reported on 31-01-2020, but there was zero death recorded in January and February month. The major outbreak happened in the month of March. In this month, total cases reported was 1251 and total death was recorded as 32 [5]. This results in the public curfew on 22-03-2020 by the Government of India. After two days, the Government of India imposed 21 days of lockdown that was from 25-03-2020 to 14-04-2020. In the month of April, the situation was little serious as the total active cases were 33,050 and total death was 1074. Total new cases reported in this month were 31,801. This results in the second lockdown of 18 days starting from 15-04-2020 to 03-05-2020. India reached an alarming stage in the month of May. The total active cases of coronavirus in India reached 165,799 which was very threatening data. 133,998 new cases were reported till 29-05-2020. Total deaths reached 4706 in this month which is very disturbing for the people. This results in the third lockdown from 04-05-2020 to 17-05-2020 (14 days), followed by fourth lockdown starting from 18-05-2020 to 31-05-2020 (14 days). The Government of India has taken the key decision for imposing the lockdowns in the country. Surely, these lockdowns affect the Indian economy a lot. But life always comes first. If these lockdowns were not imposed then might be possible that these figures will be very threatening and India will be on the top of the world's list in the active cases. This paper presents data analysis and visualization of India using Python programming. Further sections will explain the used methodology of AR and ARIMA models, case studies and discussion.

✉ N. K. Garg
nk_garg@yahoo.com

[1] Department of Information Technology, Engineering College Jhalawar, Jhalawar, Rajasthan, India

[2] Department of Electrical Engineering, Engineering College Jhalawar, Jhalawar, Rajasthan, India

Machine Learning Model used for Time Series Analysis

AR and ARIMA time series methods are a univariate (single vector) time series without a trend and seasonal component, while other time series methods are multivariate with trend and seasonal component. Both the methods are the basic and easy to understand by using Python. This typically used allows the model to rapidly adjust for the sudden changes in the trends, resulting in more accurate forecasts other than various time series forecast methods. AR and ARIMA models are used as both give the accurate forecasts as compared to other models available in time series analysis. The dataset is used as a linear function of the observations at prior time steps, hence AR and ARIMA models give more precise, accurate and suitable forecast other than various methods. The following tools have been used for the analysis of AR and ARIMA models:

Machine Learning

In simple words, developers make the machine to learn and improve from 'experience.' Here, experience is the previous data based on which predictions will be made. It is an application of artificial intelligence. This paper explains the time series analysis as a machine learning model for the future predictions.

Time Series Analysis

This paper explains the ARIMA model for the time series analysis. This model works for the supervised learning in which previous data are present. There are other algorithms present for the supervised learning like linear regression, logistic regression, while time series analysis is a statistical technique that deals with the time series data or trend analysis. Time series data means that data are in a series of particular time periods or intervals. The data are considered in three types:

Time Series Data

A set of observations on the values that a variable takes at different times.

i. **Cross-sectional data:** Data of one or more variables, collected at the same point in time.

ii. **Pooled data:** A combination of time series data and cross-sectional data.

The above dataset is time series data which were set of observations on the values that a variable taken at different times.

Autoregressive Integrated Moving Average Model (ARIMA model)

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data and as such provides a simple yet powerful method for making skillful time series forecasts. It is a generalization of the simpler autoregressive moving average and adds the notion of integration. Each of these components is explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model is also called the lag order.
d: The number of times that the raw observations are differenced is also called the degree of differencing.
q: The size of the moving average window is also called the order of moving average [6, 7].

$$ARIMA(p,d,q) : X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + Z_t \tag{1}$$

where $Z_t = X_t - X_{t-1}$.

Here, X_t is the predicted number of confirmed COVID-19 new cases at t^{th} day, α_1, α_2, β_1 and β_2 are parameters where as Z_t is the residual term for t^{th} day.

A linear regression model is constructed including the specified number and type of terms, and the data are prepared by a degree of differencing in order to make it stationary, i.e., to remove trend and seasonal structures that negatively affect the regression model. A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I or MA model. Adopting an ARIMA model [8] for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

Case Study

The dataset [5] has collected and downloaded from 1.1.2020 to 31.07.2020. Table 1 shows the data of four major columns which are total cases, new cases, total deaths and new deaths. In this research paper, only new

Table 1 Data of India (Monthly)

S.No	Month	Total cases	New cases	Total deaths	New deaths
1	January	1	1	0	0
2	February	3	2	0	0
3	March	1251	1249	32	32
4	April	33,050	31,801	1074	1042
5	May	182,143	150,342	5164	4122
6	June	566,840	416,498	16,893	14,777
7	July	1,638,870	1,222,372	35,747	20,970

cases have been considered for the reference. The new cases data are dependent variable for future prediction using ARIMA seires analysis.

Here, for observations, it has taken the four series which are total cases, new cases, total deaths and new deaths in order to understand how the coronavirus [COVID19] has affected the country. It is clear from Table1 that the situation was under control in the month of January and February. Only one series has been taken, which is new cases, for the time series analysis.

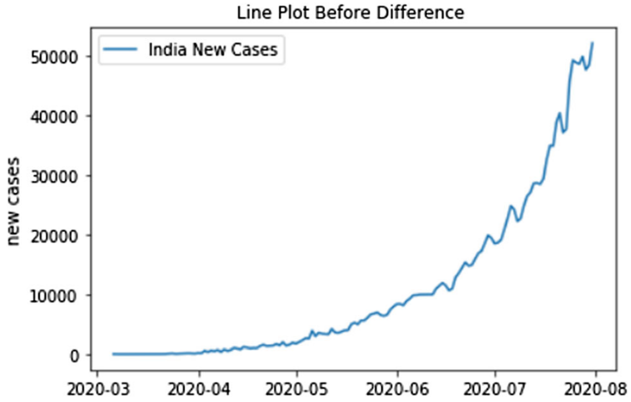


Fig. 1 Line plot of new cases in India before using diff function

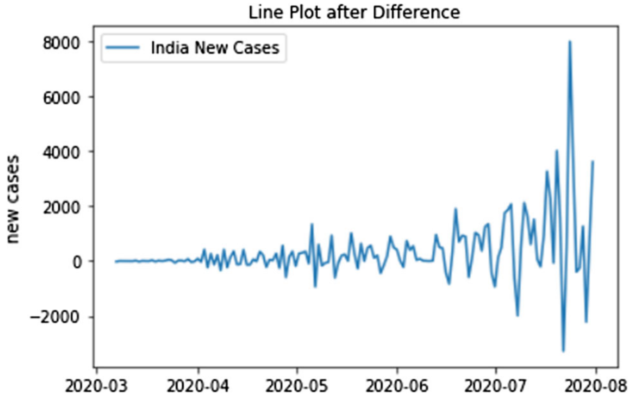


Fig. 2 Line plot of new cases in India after using diff function

Figure 1 shows that the new cases keep on increasing monthly. The average also keeps on increasing. For creating the ARIMA model, firstly, it has to make the average stationary. It can be done by using the shift function. It will shift the data by 1. It makes the 0th entry as NaN and shifted to the 1th location entry. Then, the diff function has been used which makes the difference with the period 1.

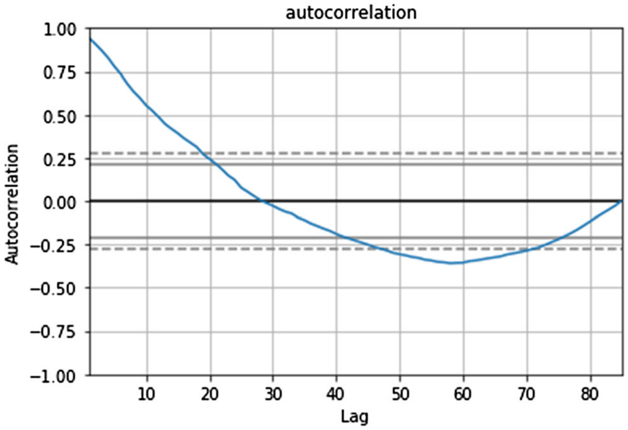


Fig. 3 Line plot of autocorrelation of new cases in India

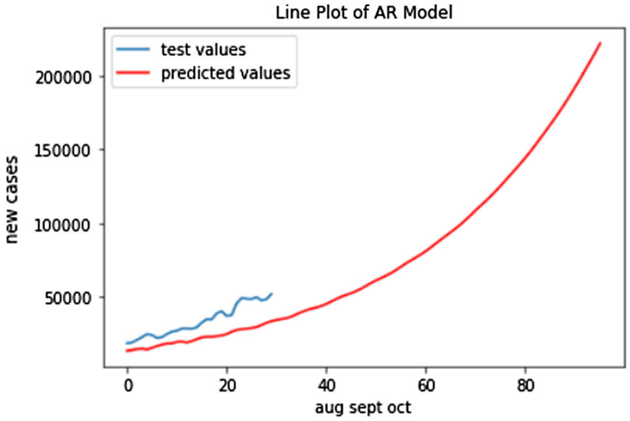
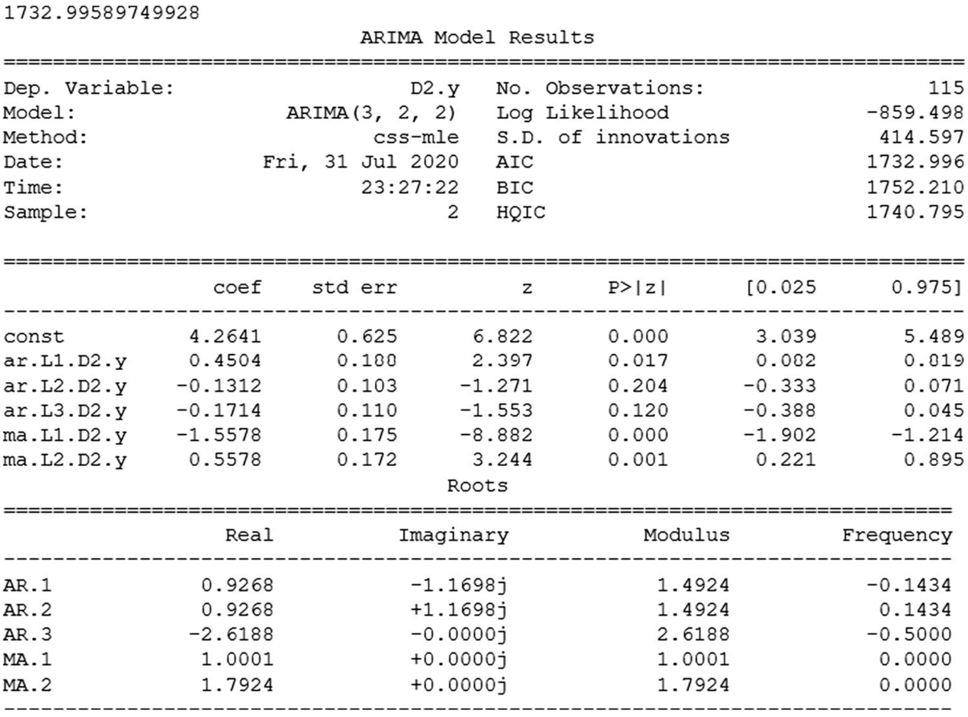


Fig. 4 Line plot of AR model

Fig. 5 ARIMA model results



```
1732.99589749928
                             ARIMA Model Results
==============================================================================
Dep. Variable:                   D2.y   No. Observations:                  115
Model:                 ARIMA(3, 2, 2)   Log Likelihood                -859.498
Method:                       css-mle   S.D. of innovations            414.597
Date:                Fri, 31 Jul 2020   AIC                           1732.996
Time:                        23:27:22   BIC                           1752.210
Sample:                             2   HQIC                          1740.795

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2641      0.625      6.822      0.000       3.039       5.489
ar.L1.D2.y     0.4504      0.188      2.397      0.017       0.082       0.819
ar.L2.D2.y    -0.1312      0.103     -1.271      0.204      -0.333       0.071
ar.L3.D2.y    -0.1714      0.110     -1.553      0.120      -0.388       0.045
ma.L1.D2.y    -1.5578      0.175     -8.882      0.000      -1.902      -1.214
ma.L2.D2.y     0.5578      0.172      3.244      0.001       0.221       0.895
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.9268           -1.1698j            1.4924           -0.1434
AR.2            0.9268           +1.1698j            1.4924            0.1434
AR.3           -2.6188           -0.0000j            2.6188           -0.5000
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2            1.7924           +0.0000j            1.7924            0.0000
-----------------------------------------------------------------------------
```

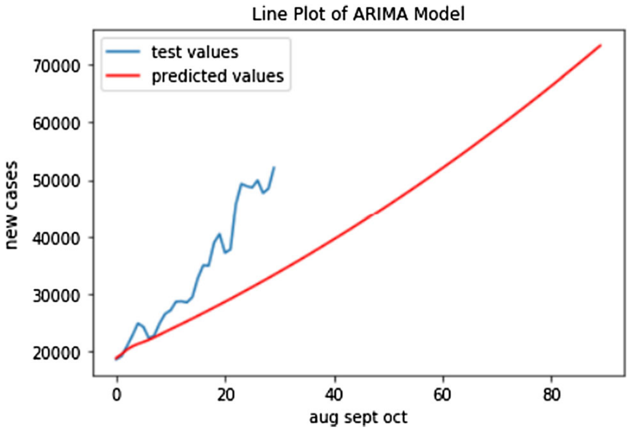


Fig. 6 Line plot of ARIMA model

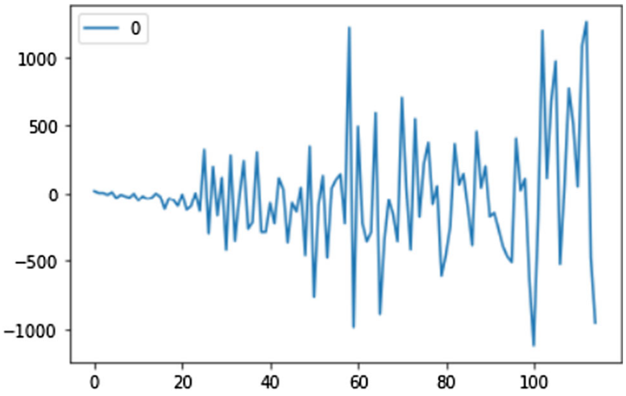


Fig. 7 Line plots of residual errors

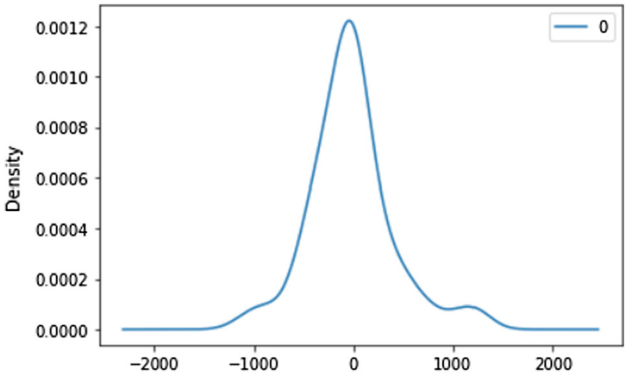


Fig. 8 Gaussian curve

This difference function is d used in the ARIMA model (*p*, *d*, *q*). After that, it has taken the data from the 1th location as the 0th location is NaN which is “not a number.”

Figs. 2and 3 show that the autocorrelation can observe that it is showing maximum four lags and it has been taken maximum four lags in ARIMA model. Now, it has been created the train set and test set from the given set of new cases in India. The total set size is 85. Out of which author has divided train set which is 80% of total dataset (which is 68) and 20% of total dataset comes under test dataset (which is 17), which can be done by using following command. In the next step, the AR (autoregression) model has been implemented. This is the first parameter of the ARIMA model which is p. Then, observe the prediction accuracy in the AR model. The line plot has been drawn of observed values and predicted values.

Table 2 Comparison between AR and ARIMA model

S.No	Date	New cases (predicted value) using AR model	New cases (predicted value) using ARIMA model
1	01-08-2020	19,127.84654	18,811.7771
2	02-08-2020	20,070.83923	19,496.73749
3	03-08-2020	21,407.63072	20,286.39556
4	04-08-2020	22,644.89493	20,864.69016
5	05-08-2020	23,094.39599	21,269.94163
6	06-08-2020	23,061.35462	21,610.66808
7	07-08-2020	23,474.98452	21,984.90175
8	08-08-2020	23,932.3733	22,415.99037
9	09-08-2020	24,834.62138	22,882.98607
10	10-08-2020	26,461.46398	23,356.58538
11	11-08-2020	27,687.27695	23,822.3357
12	12-08-2020	28,167.4772	24,281.16295
13	13-08-2020	28,546.71779	24,740.40361
14	14-08-2020	28,961.56956	25,205.71816
15	15-08-2020	29,646.71742	25,678.5349
16	16-08-2020	31,028.00423	26,157.49683
17	17-08-2020	32,508.39488	26,640.83504
18	18-08-2020	33,581.37048	27,127.68597
19	19-08-2020	34,442.1743	27,618.12537
20	20-08-2020	35,064.14707	28,112.6038
21	21-08-2020	35,630.20854	28,611.46259
22	22-08-2020	36,733.28189	29,114.78299
23	23-08-2020	38,266.622	29,622.47982
24	24-08-2020	39,736.57873	30,134.44548
25	25-08-2020	41,063.81968	30,650.62875
26	26-08-2020	42,118.93891	31,171.03528
27	27-08-2020	42,905.09984	31,695.69278
28	28-08-2020	43,912.35113	32,224.62177
29	29-08-2020	45,341.1314	32,757.82689
30	30-08-2020	46,959.79303	33,295.3028
31	31-08-2020	48,650.91991	33,837.04294
32	01-09-2020	50,210.49173	34,383.04428
33	02-09-2020	51,437.52204	34,933.30724
34	03-09-2020	52,601.89967	35,487.8335
35	04-09-2020	54,017.34442	36,046.62432
36	05-09-2020	55,687.88021	36,609.67994
37	06-09-2020	57,586.57051	37,177.00004
38	07-09-2020	59,558.19656	37,748.58421
39	08-09-2020	61,314.15078	38,324.43228
40	09-09-2020	62,878.71621	38,904.54428
41	10-09-2020	64,488.19904	39,488.92031
42	11-09-2020	66,262.43364	40,077.56044
43	12-09-2020	68,294.81872	40,670.46469
44	13-09-2020	70,560.30304	41,267.63304
45	14-09-2020	72,807.48475	41,869.06546
46	15-09-2020	74,896.70231	42,474.76194
47	16-09-2020	76,910.71904	43,084.72249
48	17-09-2020	8956.443795	43,698.94712

Table 2 continued

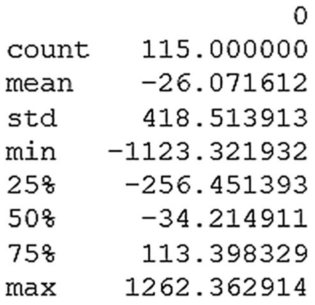
S.No	Date	New cases (predicted value) using AR model	New cases (predicted value) using ARIMA model
49	18-09-2020	81,179.17785	44,317.43582
50	19-09-2020	83,680.41134	44,940.18859
51	20-09-2020	86,334.74148	45,567.20544
52	21-09-2020	88,964.92022	46,198.48637
53	22-09-2020	91,523.73613	46,834.03137
54	23-09-2020	94,031.41345	47,473.84044
55	24-09-2020	96,598.48099	48,117.91359
56	25-09-2020	99,386.75931	48,766.25081
57	26-09-2020	102,408.2989	52,071.89803
58	27-09-2020	105,546.7145	52,745.8197
59	28-09-2020	108,705.1312	53,424.00544
60	29-09-2020	11,820.82959	54,106.45526
61	30-09-2020	114,918.9908	54,793.16915
62	01-10-2020	118,138.9161	55,484.14711
63	02-10-2020	121,577.9106	56,179.38915
64	03-10-2020	125,216.5696	56,878.89527
65	04-10-2020	128,986.2917	57,582.66545
66	05-10-2020	132,793.8538	58,290.69971
67	06-10-2020	136,584.4184	59,002.99805
68	07-10-2020	140,422.3901	59,719.56046
69	08-10-2020	144,416.1772	60,440.38695
70	09-10-2020	148,619.4509	61,165.4775
71	10-10-2020	153,028.351	61,894.83214
72	11-10-2020	157,576.9934	62,628.45084
73	12-10-2020	162,175.9382	63,366.33363
74	13-10-2020	166,811.1367	64,108.48048
75	14-10-2020	171,546.9959	64,854.89141
76	15-10-2020	176,457.5911	65,605.56642
77	16-10-2020	181,593.367	66,360.5055
78	17-10-2020	186,945.6771	67,119.70865
79	18-10-2020	192,442.4332	67,883.17588
80	19-10-2020	198,025.497	68,650.90718
81	20-10-2020	203,699.9736	69,422.90255
82	21-10-2020	209,514.5733	70,199.162
83	22-10-2020	215,536.6505	70,979.68553
84	23-10-2020	221,808.0195	71,764.47313
85	24-10-2020	228,304.4721	72,553.5248
86	25-10-2020	234,969.0419	73,346.84055

Fig. 4 shows the line plot of observed values and the predicted values using AR model. It can be observed that it requires finer tuning between observed and predicted values. Finally, ARIMA model has been implemented in order to observe the relation between observed values this is not but the test data and the predicted values. As per the above discussion, ARIMA model requires three parameters which are *p*, *d*, *q*. After that it can be seen the summary of the ARIMA model result and implementing akaike information criteria function using Python. It is a single number score that can be used to determine which one of the multiple models is most likely the best model for the given set of combinations. Fig. 5 shows the ARIMA model summary and the aic value. This is the best value in the multiple combinations of (*p*, *d*, *q*). This aic value is achieved by (2, 2, 4) combination.

This is the test data and the predicted data using ARIMA model, the prediction values are 17 as the step = 17. Now we observe the line plot (Fig. 6) of the above result.

As it can be observed that now the relation between observed values and predicted values is finely tuned with the combination of (p = 2, d = 2, q = 4). Now it can be shown that the residual plot of ARIMA model by using resid function. Line plot (Fig. 7) of the residual errors, is shown that there may still be some trend information not captured by the model.

Next, it is defined by a density plot (Fig. 8) of the residual error values, suggesting the errors are Gaussian but may not be centered on zero.



```
                    0
count      115.000000
mean       -26.071612
std        418.513913
min      -1123.321932
25%       -256.451393
50%        -34.214911
75%        113.398329
max       1262.362914
```

The distribution of the residual errors is displayed. The results show that indeed there is a bias in the prediction (a nonzero mean in the residuals). Note that, although above it is used the entire dataset for time series analysis, ideally it would perform this analysis on just the training dataset when developing a predictive model. Table 2 shows that data are more precise in ARIMA model as compared to AR model.

Conclusion

From the above results and line plots, it is shown that the predictive values from ARIMA model are more fit than AR model. These predictive values are for next coming days. It is clear that the data of new cases will keep on increasing as the predictive values keep on increasing. In the coming future, human being has to fight with coronavirus and follow the Government of India guidelines. Lockdown is not the only solution. Lockdown is not the only solution, but it is also the social and moral responsibility of each and every citizen to follow the guidelines. Today the whole world is at high risk. In India most of the active cases are asymptomatic which mean that person is infected but does not show the symptoms of the coronavirus. The name of the disease is COVID-19, which is dangerous. People have to be well aware of the virus. Presently, India's new cases data are not so alarming as compared to our population. This is only possible with the series of lockdowns. New cases data will keep on increasing for some time after unlocking the country, but after some time, these data will be under control.

References

1. Chih-Cheng Lai, Cheng-Yi Wang, Ya-Hui Wang, Po-Ren Hsueh, Global coronavirus disease 2019: What has daily cumulative index taught us? Int J Antimicrob Agents **55**(6), 106001 (2020). https://doi.org/10.1016/j.ijantimicag.2020.106001
2. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. Lancet **395**, 497–506 (2020)
3. https://covid19.who.int/?gclid=Cj0KCQjw_ez2BRCyARIsAJfgksiJkE56RN9BAqkKycd3q–lzP_4Tq7DJjZTf02A2ZPRWZsvfCl0tcaAh-OEALw_wcB Accessed on 31.7.2020
4. L. Wang, Y. Wang, D. Ye, Q. Liu, Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. Int J Antimicrob Agents. **55**(6), 105948 (2020). https://doi.org/10.1016/j.ijantimicag.2020.105948 (Erratum in: Int J Antimicrob Agents. **56**(3), 106137 (2020))
5. https://github.com/owid/covid-19-data/tree/master/public/data
6. https://www.mathsisfun.com/data/standard-deviation-formulas.html
7. S.I. Alzahrani, I.A. Aljamaan, E.A. Al-Fakih, Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J. Infect. Public Health (2020). https://doi.org/10.1016/j.jiph.2020.06.001
8. R.K. Singh, M. Rani, A.S. Bhagavathula et al., Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surveill. **6**(2), e19115 (2020). https://doi.org/10.2196/19115