# COVID-19 Analysis Report

Abhishek D Sawalkar

## Statement of Purpose:

I was unfortunate to contract COVID-19 during the second wave in India. Time-series graphs, denoting the caseload were omnipresent in this period. I found that time series analysis resonated with me since it used mathematical equations to understand and give meaning to perpetual events. Under the guidance of Professor Supratim Biswas, at IIT Bombay, for over a year, I was able to attain solutions for predicting the cases. He emphasized the importance of reading research papers in the field before starting work. We extensively reviewed the topic by maintaining a shared repository of papers and discussing them every Saturday evening. It was a great learning curve where I became aware of the approach needed while reviewing the literature to find research gaps and a new angle to contribute to the field. I studied various parameters such as dataset, formulae, models from the area which helped me in the comparative study. I learned that Data Science isn't only about designing and using models. It is also about the transformation of data which is as important as choosing the right model for the problem.

Data was taken from Johns Hopkins University, and the focus of the analysis was on Indian cases. The time period for data was from 1st January 2020 to 9th September 2021. The split was 80-20 between training and test data. I performed data stationarity tests such as plotting the rolling mean, standard deviation, and Augmented Dickey-Fuller test and found that the dataset's stationarity could be better. Hence, I applied differencing to the dataset with a lag factor of 1.

## Project Flow:

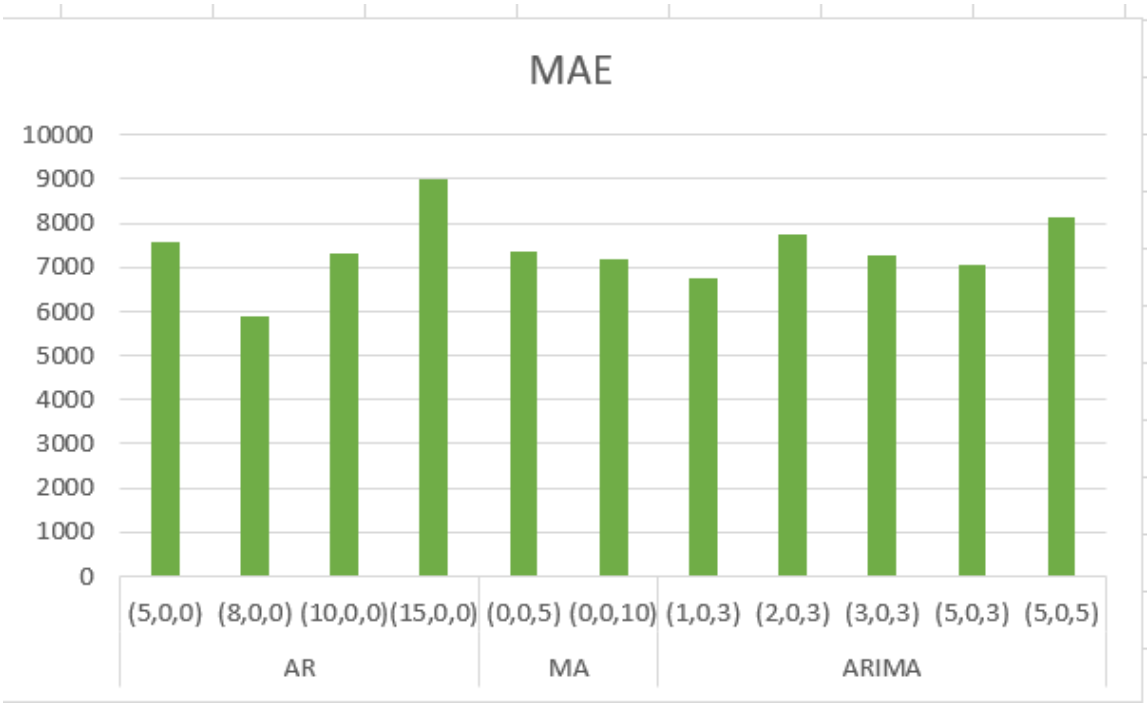
1. Get data - CSV file - Johns Hopkins University GitHub
2. Convert it into dataframe
3. Calculate the number of daily cases column
4. Apply stationarity tests
    a. Visual Graph-Rolling Mean, Standard Deviation
    b. Augmented Dickey Fuller Test
5. Check for seasonality
6. Apply transformation to change series into a stationary series
7. Divide data into train and test
8. Fit model on train data
9. Make predictions on test data
10. Graph the predicted values along with the test values
11. Calculate Accuracy Parameters
    a. Mean Absolute Error
    b. Mean Square Error
    c. Root Mean Square Error
12. Plot other graphs
    a. Error Residual Graphs

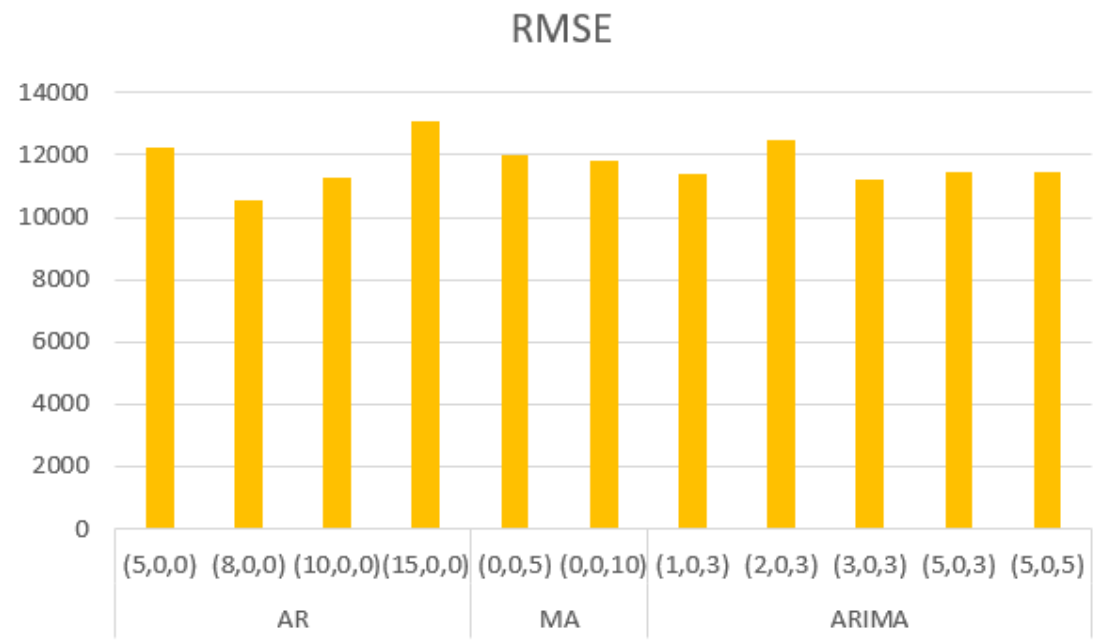
b. Value Density Plot

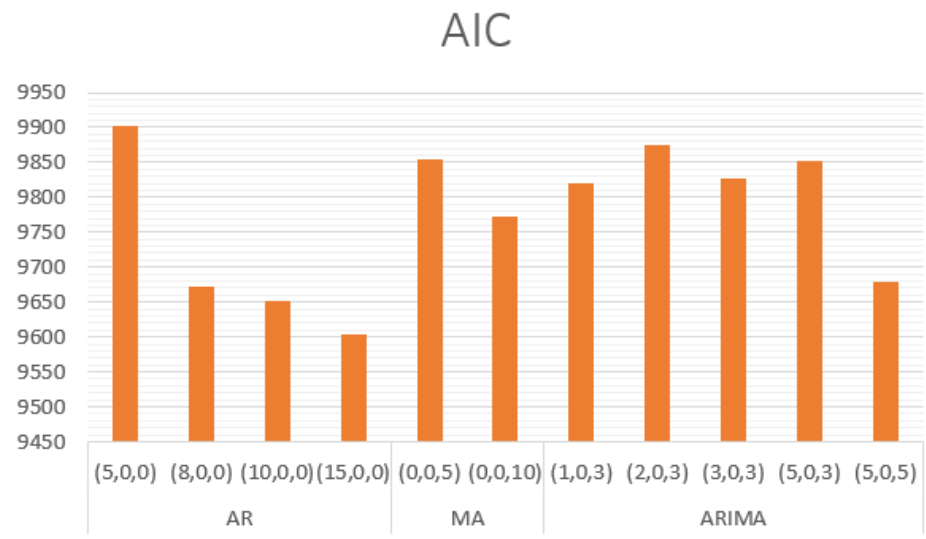
## Results

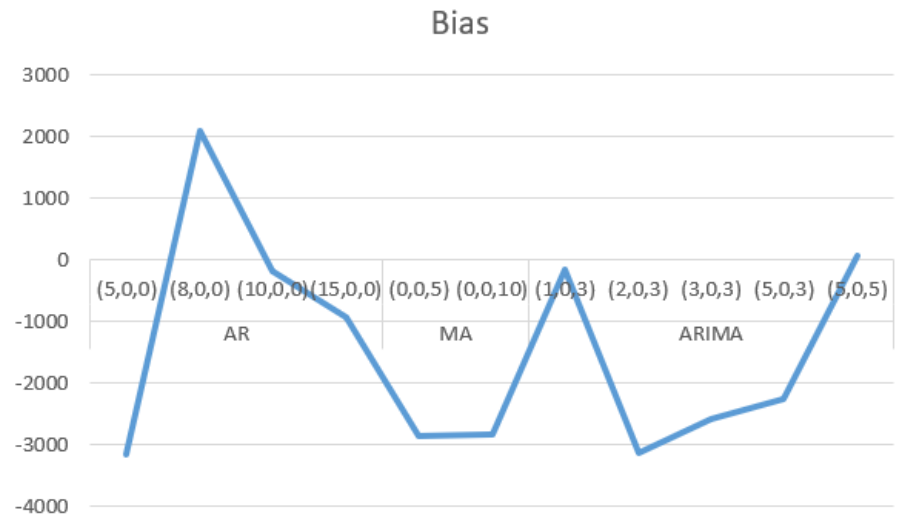
Initial analysis with time series forecasting models AR (i.e., Auto-Regressive) and MA (i.e., Moving Average) resulted in a mean absolute error of about 7440 and 7410 cases with an accuracy of 75%. Hence, I proceeded with the ARIMA (AR Integrated MA) model which used differencing and took the pithy features of both AR (predicting values based upon previous values) and MA (correcting values according to previous errors). This resulted in a boost in the accuracy of the predictions to 80% and reduction in the mean absolute error to 7396.

| **Model** | **Order** | **Coefficients** | **AIC** | **Bias** | **MAE** | **RMSE** | **Average MAE** |
|---|---|---|---|---|---|---|---|
| | (5,0,0) | 3 | 9902 | -3156 | 7559.6 | 12249 | |
| | (8,0,0) | 1,2,3,6,7,8 | 9671.9 | 2099 | 5886.5 | 10571 | |
| | (10,0,0) | 1,3,4,6,7,8,9,10 | 9650.9 | -168.8 | 7325 | 11255 | |
| AR | (15,0,0) | 1,3-13 | 9604 | -925.3 | 8989.9 | 13097 | 7440.2525 |
| | (0,0,5) | 1,2,4,5 | 9853.7 | -2854 | 7376.7 | 11995 | |
| MA | (0,0,10) | 1,3,6-9 | 9771.7 | -2826 | 7173.2 | 11804 | 7410.925 |
| | (1,0,3) | 1-5,1-5 | 9821.1 | -157 | 6753.4 | 11402 | |
| | (2,0,3) | 1-3,1-3 | 9875 | -3126 | 7736.8 | 12456 | |
| | (3,0,3) | 1, 1,3 | 9827.4 | -2581 | 7293.8 | 11199 | |
| | (5,0,3) | 1-2,1-3 | 9851.1 | -2259 | 7073.5 | 11444 | |
| ARIMA | (5,0,5) | 1-4,1-3 | 9677.9 | 73.17 | 8123 | 11465 | 7396.092 |


MAE
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
(5,0,0)
(8,0,0)
(10,0,0)
(15,0,0)
(0,0,5)
(0,0,10)
(1,0,3)
(2,0,3)
(3,0,3)
(5,0,3)
(5,0,5)
AR
MA
ARIMA


RMSE
14000
12000
10000
8000
6000
4000
2000
0
(5,0,0)
(8,0,0)
(10,0,0)
(15,0,0)
(0,0,5)
(0,0,10)
(1,0,3)
(2,0,3)
(3,0,3)
(5,0,3)
(5,0,5)
AR
MA
ARIMA




## Aim of Literature Review:

To study multiple research papers on time series analysis methods applies towards COVID-19 to gain enough technical knowledge on the topic in order to develop a comparative technical report and also perform some practical experiments with relatively recent data

## Comparative Study:

| Sr. No. | Paper Details | |
|---|---|---|
| 1 | Paper Name | Predicting the New Cases of Coronavirus [COVID-19] in India by Using Time Series Analysis as Machine Learning Model in Python – Institution of Engineers (India) |

| | | |
|---|---|---|
| | Brief Outline | • The paper uses mainly 2 models for time series prediction. ARIMA and AR.<br>• The time series is not stationary since the mean is not constant. A differencing factor of 1 is used to overcome this.<br>• The main work done is of prediction on 20 % test data set of daily new cases. ARIMA model comes ahead of AR model since it contains both difference and moving average factors. |
| | Dataset | • Taken from OWID. Data about India<br>• The dataset used is of 7 months from 1 Jan to 31 July 2020.<br>• Only 1 dependent variable – confirmed cases<br>• Differencing applied to make data stationary |
| | Formulae | $ARIMA(p,d,q) : X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + Z_t \quad (1)$<br>where $Z_t = X_t - X_{t-1}$. |
| | What more could be done? | • Analysis can be done on more amount of data, which may contain seasonality<br>• Other time series analysis methods could be applied to present comparative results<br>• Analysis can be done on more variables like recovered, deaths |
| 2 | Paper Name | Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET – 5th International Conference on Computer Science and Computational Intelligence 2020 |
| | Brief Outline | • The aim of the paper is to test if ARIMA and PROPHET can perform accurate predictions on time series data which is having a trend i.e. doesn't have a constant mean and is non seasonal.<br>• ARIMA required some steps –<br>i. Checking if data was stationary<br>ii. Transforming data to stationary<br>iii. Finding out parameters by ACF and PACF graphs<br>iv. Predicting data after fitting model<br>• PROPHET was pretty straightforward in the sense that none of the above steps were required. Just forecasting after fitting.<br>• Result - Both models wane in accuracy after some time interval. But PROPHET performs better since the beginning while ARIMA underperforms throughout. PROPHET also comes ahead in terms of accuracy for confirmed cases. Both models also have positive bias |
| | Dataset | • Taken from Kaggle. Exact source not given. Data about Indonesia<br>• Total - 1 Jan to 21 May 2020. Test Data – 22 April to 21 May 2020<br>• The data consists of 3 dependent variables - confirmed, deaths, recovered. |

| | | |
|---|---|---|
| | | • Positive Trend seen in data – non stationary<br>• Log scaling and differencing applied to make data stationary |
| | Formulae | Arima Model<br>$y't=c+\phi 1y't-1+\cdots+\phi py't-p+\theta 1\varepsilon t-1+\cdots+\theta q\varepsilon t-q+\varepsilon t,$ |
| | What more could be done? | • Analysis could be done on more data, since analysis is done by training only on 4 months of data<br>• More methods could be used for comparison |
| 3 | Paper Name | Time Series Analysis of COVID-19 Data to Study the Effect of Lockdown and Unlock in India. Saswat et.al. |
| | Brief Outline | The aim of the paper is to compare different time series models and use the best one to forecast the positive cases of COVID19 for the month of August. The compared models were ARIMA, PROPHET, TBATS, N-BEATS, Moving Average, Single Exponential, Double Exponential. It was found that ARIMA had the least RMSE from all the models.<br>Dataset - Positive Trend with no seasonality. Log Transformation and Differencing used to make data stationary<br>ARIMA (5,1,4) was used for lockdown period<br>ARIMA (1,1,1) was used for unlock period<br><br>Given 2 scenarios i.e. of lock down or unlock. It was found that the slope of positive cases gets steeper for unlock period while continuing the same trend as it had shown in lock down period. A new model was also proposed which also factored in the number of tests w.r.t. the positive cases and it showed that with increase in number of tests the positive cases also increase with a positivity rate of 7-11%. |
| | Dataset | Source: PR legislative research India<br>Date: 12 March to 2 July 2020<br>Lockdown Data: 12 March to 7 June<br>Unlock Data: 8 June to 2 July<br>Prediction Made for: August<br>Dependent Variables: 3 i.e. Confirmed, Death, Recovered. Only Confirmed cases taken for analysis |
| | Formulae | Data Transformation: $Y_t = log(X); \quad Y_t = Y_t - Y_{t-m};$<br>ARIMA: $Y_t = \omega + \phi \times Y_{t-1} + e_t$ |

| | | |
|---|---|---|
| | What more could be done? | • Analysis could be done on latest data, since analysis was done only on 5 months of data.<br>• Creation of ensembles of more models with multivariate analysis and more direct and indirect factors<br>• Using transfer learning to bring learnings of data of one country to any other |
| 4 | Paper Name | Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. Tandon et.al. |
| | Brief Outline | The aim of the paper is to forecast covid cases for 14 April to 3 May after comparative study of some models.<br>A comparison between the positive cases between India and most infectious countries of the world is also done along with a specific comparison between India and other South Asian countries.<br>ARIMA model (2,2,2) was the one with the least error and its residual error graph follows normal distribution suggesting that the standard deviation of the dataset is constant. |
| | Dataset | Source: John Hopkins University. Mainly Focussed on Positive cases in India<br>Test Data: 22 Jan to 13 April 2020<br>Train Data: 14 April to 3 May 2020 |
| | Formulae | $ARIMA(p,d,f): X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + Z_t$<br>$Z_t = X_t - X_{t-1}$ |
| | What more could be done? | Analysis could be done on latest data, since analysis was done only on 5 months of data.<br>The paper doesn't do an analysis of the data of other countries. The cases have just been plotted |