

Short-Term Forecasting of India's Corona Virus Outbreak Using a Hybrid Modeling Approach

Mehuli Paul¹ and Meghanto Majumder²

[Received on May, 2020. Accepted on March, 2021]

ABSTRACT

The first few cases of the novel coronavirus outbreak can be traced back to those that occurred in the Chinese city of Wuhan in December, 2019. On 30th January, 2020, the outbreak was declared as a Public Health Emergency of International Concern and later, on 11th March, 2020, it was declared to be a pandemic. As of 24th May, 2020, it spreaded over 188 countries and territories and the total number of cases stood out at over 5.4 million, with over 346,427 deaths. In this paper, we restrict our attention to the outbreak in India. The first positive case was reported on 30th January, 2020 as per records. Number of confirmed cases (in India) stands at 138,535 with 4023 deaths as of 24th May. The main focus of our paper is to propose a Hybrid ARIMA (Auto Regressive Integrated Moving Average) model with error remodeling using Fourier Analysis performed on the number of daily new cases. Our aim is to show how the hybrid model generates better short-term forecasts compared to single ARIMA model and also how the modeling on a data set of the first phase of the lockdown generates more accuracy in the forecasting. These short-term forecasts for the number of daily new cases can guide us and throw some light on the growth pattern of COVID-19, thus guiding the government to make arrangements accordingly.

1. Introduction

The world was caught unaware by the COVID-19 pandemic in December 2019, when Wuhan, the city of China experienced its first case of coronavirus. It quickly spread to the other parts of the world, infecting almost 188 countries and territories and causing deaths in large numbers. It was declared to be a pandemic on 11th March, 2020 by World Health Organization (WHO, 11th March, 2020).

✉ : Mehuli Paul
E-mail: mehuli582@gmail.com.

The world is struggling to find a solution to this pandemic, with over 5.4 million confirmed cases and 346,427 deaths as of 24th May, 2020 (ArcGIS, 24th May 2020). The primary purpose of the paper is to focus on the current situation in India. As per sources, the total number of cases have crossed 138,535 and deaths have surpassed 4023 in India as of 24th May (<https://www.mohfw.gov.in>, 24th May, 2020). The disease was declared to be pandemic by WHO as it became more widespread and infected men. First phase of lockdown was announced on 24th March (ndtv.com, 11th April, 2020). Nation wide lockdown from 25th March. It motivates us to do some research on COVID-19.

For this we make use of a Hybrid Auto Regressive Integrated Moving Average (or ARIMA) model with error remodeling using Fast Fourier Transform (or Hybrid ARIMA-FFT) and notice that it is effective in making short-term forecasts. Many other short-term forecasting methods using the individual models have been proposed earlier for infectious diseases such as dengue fever and malaria, etc. We first forecast using data from 30th January, i.e. the day when the first case was confirmed in India and then perform similar study using the data starting from the first phase of nation-wide lockdown, i.e. data starting 25th March. We have also shown how hybrid model gives more accurate forecasts compared to single ARIMA and it can also be seen that how more accurate forecasts can be generated by fitting the model to the data starting from first phase of lockdown. This is perhaps owing to the fact that though India recorded its first positive case on 30th January, the spread of the disease was slower earlier with zero confirmed cases in the following days. The spread of the disease gained momentum only after a certain period, and hence called for a nation-wide the lockdown at that time.

2. Forecasting the COVID-19 Outbreak

We focus on the daily number of confirmed cases in our country, India. We take the data from api.covid19india.org/csv, a crowd sourced website that obtains their data from various central and state government websites. The datasets constitute the number of confirmed cases on a daily basis, collected starting from the onset of the disease in India, i.e., from 30th January, 2020. We take one dataset with observations starting 30th January and another with those starting 25th March. We use these datasets to produce forecasting using our proposed Hybrid ARIMA-FFT model. All datasets and codes used are available at <https://github.com/mehulipaul/covid-india>.

a) Datasets

Datasets on daily number of confirmed cases, starting from 30th January, 2020 to 24th May, 2020 have been downloaded from api.covid19india.org/csv. It was then

made univariate by editing, i.e., we kept only the data of daily number of confirmed cases. It has a total of 117 observations and we take the first 110 observations as our training dataset. The remaining 7 observations from our testing dataset (since our goal is to generate short-term forecasts, we restrict the forecasting period to 7 days). We have used two datasets for both our approaches – one that constitutes all observations starting 30th January, the day of onset of the disease in India and another constituting of datasets starting 25th March, the day when the nation-wide lockdown was imposed. Let us call the former ‘Datatype-1’ and the latter ‘Datatype-2’ for our convenience. This has been done since the initial spread of the disease was slow and intermittent. Also, not enough awareness or security measures were taken to prevent the spread of the disease. We take the last seven observations for both the datasets as our testing data and the rest is taken as training data.

The table below, Table 1 in the next page shows the trends of the training datasets for both Datatype-1 and Datatype-2 and their ACF (Auto correlation function) and PACF (Partial autocorrelation function) plots. Since no seasonality is observed in the datasets, we do not take seasonal models into consideration.

Training datasets for both data types and their ACF, and PACF plots

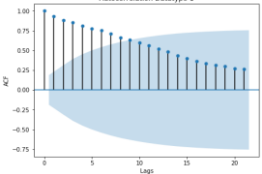
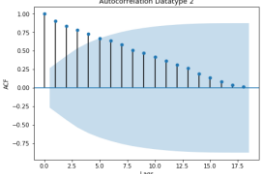
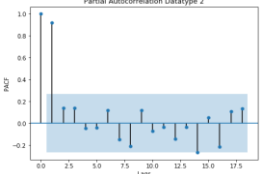
Datatype	Training Data	ACF Plot	PACF Plot
Datatype-1			
Datatype-2			

Table 1

b) Proposed Model

Here we apply a hybrid modeling approach for forecasting. We propose a hybrid model combining ARIMA and Fast Fourier Transform (FFT) techniques together so as to forecast the outbreak better and reduce the errors generated by single

time series models. Let us now discuss the individual models, along with the proposed hybrid model in brief.

ARIMA Model

The ARIMA model is a linear time series model used in statistics and econometrics and can be expressed as–

$$y_t = \theta_0 + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + \dots + \Phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - \dots - \theta_q \varepsilon_{t-q} \tag{2.1}$$

Non-seasonal ARIMA model is usually denoted as ARIMA (p, d,q), where p, d and q are the three parameters denoting the number of lag(order of autoregressive or AR model), the degree of differencing and the order of moving-average In the above equation, the actual value of the concerned variable at any time t (here, number of confirmed cases) is denoted by y_t , and ε_t denotes the random error at time t, where θ_i ($i = 1, 2, \dots, q$) and Φ_j ($j = 1, 2, \dots, p$) are the associates with the past values of observations and past values of error terms, respectively. The model assumes that the set of errors are independent and identically distributed random variables. The chosen parameters depend on the ACF (Auto-correlation function) plot, PACF (Partial autocorrelation function), AIC (Akaike Information Criteria) value, and BIC (Bayesian Information Criteria) value (Hyndman, 2018; Box, 2008).

Fast Fourier Transformation

Fast Fourier Transformation is an algorithm used to compute the discrete Fourier Transformation (DFT) of a sequence. It converts signals from its original domain of time or space into frequency domain by decomposing the original signal into a sum of sinusoids of different frequencies, amplitudes and phases (Heideman, 1984). It reduces the complexity of calculating DFT and the difference in speed is enormous, especially for longer datasets. FFT was described as ‘the most important algorithm of our lifetime’ by Gilbert Strang in 1994 (Strang; Kent, 1992).

Mathematically, DFT is represented as:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{i2\pi kn}{N}} \tag{2.2}$$

where, X_0, \dots, X_{N-1} are ($K=0,1, \dots, N-1$) and $e^{\frac{i2\pi}{N}}$ is a primitive Nth root of 1. Evaluation of this requires $O(N^2)$ operations: there are N outputs X_k and each output requires a sum of N terms. FFT computes the same results in $O(N \log N)$ operations and all known FFT algorithms require $O(N \log N)$ operations, although we cannot dismiss the chances of the possibility of a lower complexity score (Frigo, 2006).

The Hybrid Model

Proposed hybrid model is constructed by combining ARIMA with error-remodelling using FFT in a two-step approach. We start by applying ARIMA to both Datatype-1 and Datatype-2 and computing its residues, or the error dataset which is seen to be non-stationary and oscillatory. We remove any linear trend found from residuals and model the respective de-trended residuals using FFT. Fourier Transform converts the residuals from its original time domains to frequency domains. We perform the inverse FFT to compute the original Fourier approximation of the de-trended error dataset. We re-apply the eliminated trend to the dataset to obtain our final predictions for the error dataset, as well as get an m-step forecast of the same (Fumi, 2013). The final prediction is thus given by:

$$\hat{y}_t = y_t + z_t \quad (2.3)$$

where, \hat{y}_t is the final prediction (generated by the hybrid model), y_t and z_t are the predictions generated by ARIMA and FFT respectively.

The analysis is performed using statistical packages in Python 3 (Miller, 2019; McKinney, 2011; Seabold, 2010). We have presented the step-by-step approach as an algorithm.

Algorithm for the Hybrid ARIMA-FFT Model

- i) Input the training data for Datatype-1 and Datatype-2 separately.
- ii) Find the best-fit ARIMA (p, d, q) model for each dataset. Parameters p, d and q are chosen as per the method mentioned in 2.2.1
- iii) Obtain predictions generated by ARIMA (p, d, q) fitted to each training dataset and also find the required m-step ahead (m is an integer) forecast for each data type using testing data.
- iv) Obtain residuals (ϵ_t) for each model by subtracting the predicted values from the actual ones.
- v) Obtain a linear trend if found in the residual data, and subtract it from the data to de-trend it.
- vi) Choose the number of wave components to be considered, k as proportional to the logarithm of the dataset size, n, as observed to be the best fit.
- vii) Calculate the FFT on the de-trended data, and create the frequency spectrum.
- viii) Except for f_0 , order f_1 to $f_{n/2}$ components by decreasing order of amplitude.
- ix) Perform the inverse FFT on the dataset, taking into account only the first k components in the ordered frequency spectrum.

- x) Re-apply the subtracted trend found in step 5 to generate the fitted residuals.
- xi) Since the result of inverse FFT is periodic in nature, extract the first m values, and add the trend constant for n to get the m -step forecasts of the residuals.
- xii) Now, add the forecasts generated in both step 3 and step 7 to obtain the final forecast (\hat{y}_t) for each data type.

3. Results

In this approach, we deal two datasets as explained above – one that starts from 30th January, i.e. the day of the first confirmed case in India or Datatype-1 and second, the one that starts from 25th March, 2020 the start of the nation-wide lockdown in India or Datatype-2. For Datatype-1, we first divide the dataset into training data (taking first 110 observations into consideration, i.e. from 30th January to 10th May) and testing data (taking the last 7 observations into consideration, i.e. 11th May to 24th May). The pattern and ACF, PACF plots for this are already shown in Table 1. They can help us determine the order of the models. Unit root tests are also performed and all data points are seen to be non-stationary. We choose the ARIMA order based on a stepwise search to find an ARIMA model with the lowest AIC and BIC values. Let us, for the purpose of our convenience, name the two hybrid models fitted to Datatype-1 and Datatype-2 as Hybrid-1 and Hybrid-2 respectively. Figure 1 in the next page shows plots of ARIMA residuals.

Plots of ARIMA residuals for a) Datatype-1 b) Datatype-2.

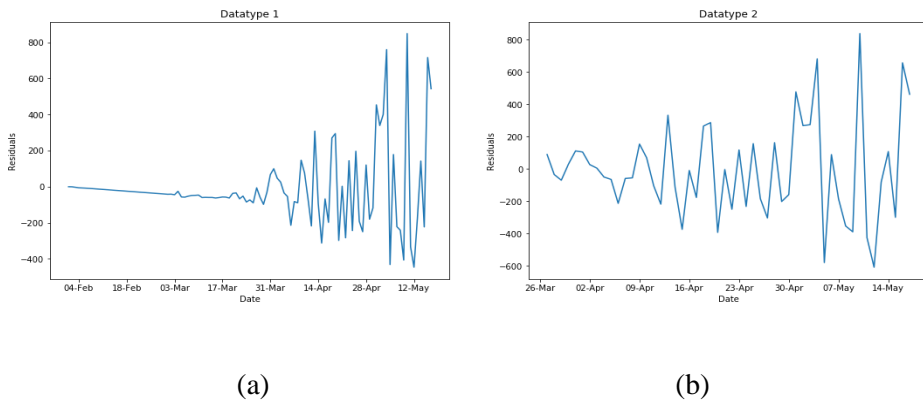


Figure 1

Hybrid 1: We proceed with Datatype-1. The best-fit ARIMA for Datatype-1 is found to have the order (2, 2, 3). We compute the residuals and generate the predictions for the same. Now, we perform seven steps ahead forecast using the testing data and compare this with the actual observations. The residuals are found to be oscillatory and are remodeled using FFT into the frequency domain. We choose an appropriate number of frequency components ordered by amplitude to generate the Fourier approximation of the de-trended residual function. We convert the function back to the time domain, and re-apply the found trend to get our in-sample predictions. As the Fourier Transform and its inverse return a periodic function, thus m-step forecasts can be generated by repeating the initial data signal for m steps with the correct trend applied. We add these to those generated by ARIMA (2, 2, 3). This addition gives us the final hybrid model predictions and forecasts. Predictions for each model are shown in Figure 2.

Hybrid 2: We proceed with the Datatype-2. We approach in a similar manner by first dividing this into training (from 25th March, 2020 to 10th May, 2020) and testing datasets (from 11th May, 2020 to 24th May, 2020). The best-fit ARIMA is found to have the order (1, 2, 1). We compute the residuals and generate the predictions for the same. Seven steps ahead forecasts are generated using the testing data. The computed residuals are found to be oscillatory and are remodeled using FFT. We choose an appropriate number of frequency components ordered by amplitude to generate the Fourier approximation of the de-trended residual function. We convert the function back to the time domain, and re-apply the found trend to get our in-sample predictions, similar to the operations performed in Hybrid 1. We generate forecasts for the residuals modeled in the Fourier approximation and add these to those generated by ARIMA (1, 2, 1). This addition gives us the final hybrid model predictions and forecasts. Predictions for each model are shown in Figure 2.

Depicts the plots for (a) Actual vs. Predicted for the fitted model, Hybrid-1
 (b) Actual vs. Predicted for the fitted model, Hybrid-2.

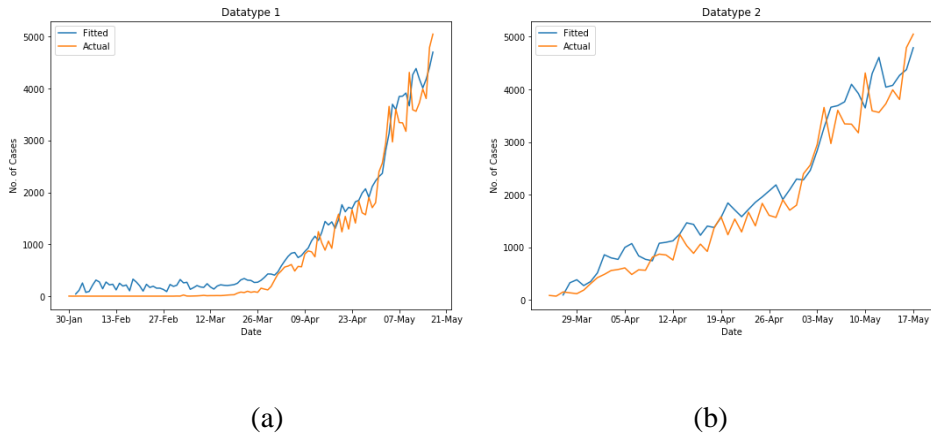


Figure 2

We have shown the respective forecasts generated by both single ARIMA and the Hybrid ARIMA-FFT models for each dataset in Figure 3 below.

Forecasts generated by ARIMA and Hybrid ARIMA-FFT models for Datatype-1 and Datatype-2

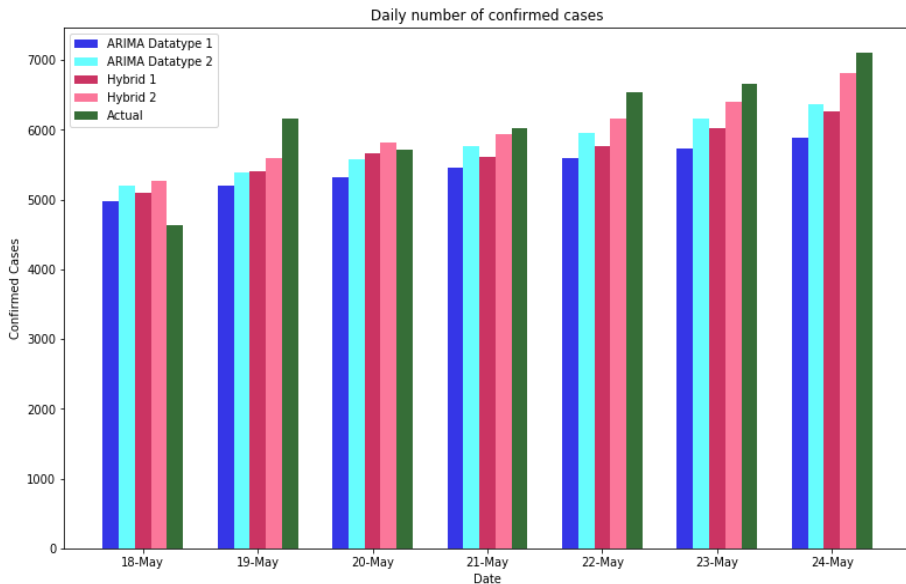


Figure 3

To judge and compare the performances of both the single ARIMA and the Hybrid ARIMA-FFT fitted separately to the two types of datasets, Datatype-1 and Datatype-2, we have computed their respective error percentages and presented it in the form of a bar graph in Figure 4 for better understanding. To further judge the effectiveness of the models in forecasting the rise in daily number of cases for both datasets, we have also computed the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) for each model fitted to each type of dataset and presented them in Table 2.

MAE and RMSE of ARIMA and the Hybrid ARIMA-FFT model

Datatype	Performance Parameters	ARIMA	Hybrid ARIMA-FFT Model
Datatype-1	RMSE	826.35	620.32
	MAE	767.39	567.04
Datatype-2	RMSE	552.89	386.18
	MAE	509.19	332.95

Table 2

From the above discussions, it is clear that the Hybrid ARIMA-FFT model is more accurate in forecasting of short-term outbreaks compared to ARIMA alone. What is also notable is that the data starting from the first phase of lockdown, i.e. Datatype-2 is adjusting better to the trend and giving superior forecasts as compared to Datatype-1. The Hybrid model, on Data type 2 has the predictions with the least RMSE and MAE values.

4. Limitations

In India, according to GISAIID, three viral clades (or strains) of coronavirus have so far been traced – commonly found in Iran, Italy and Wuhan (China), respectively. It is suspected that people who travelling to India from other countries in the recent months may bring other strains of the virus. Also, as per “no free lunch theorem”, no model is optimal everywhere or in every situation (Wolpert, 1997). We do not consider how the performance of the model changes if more strains are to be found in future time. This along with various other

controlling factors may determine the actual rise in the number of cases and its pattern in the future. The results may differ if the same models are applied to datasets of a different country. Since limited data is available, the performance proposed models strongly depend on the degree of availability of data. The ARIMA being a linear time series is efficient in capturing the linear trends only.

5. Discussions

Almost all nations are in the grip of the pandemic. The sudden outbreak of such an infectious disease without any prior warning has thrown a challenge to the modelers all over the world. Modeling the outbreak perfectly is very difficult owing to the nature of the virus which mutates quite fast. Also, another challenge is that different mutations are present in different parts of the world. All these make the building of a perfect model which might be applicable to all countries or the same country but in all situations next to impossible. We have tried to present a hybrid modeling approach using ARIMA and Fourier analysis to capture both the linear and non-linear points in the datasets. Also, we suggest using the Datatype-2 or the data starting from the day of first phase of lockdown since the cases in India gained momentum only from March and also because only limited data is available for the earlier trend (this is perhaps because not all states had taken to doing a minimum number of required tests on a daily basis and for reasons alike). We have thus presented our forecast using the hybrid modeling technique, while also pointing out how the short-term (7 steps ahead) forecasts performed have shown better accuracy when the model was built on data starting from the first phase of lockdown. This approach will be useful in depicting the short-term rises and can help the responsible institutions in taking the necessary steps.

6. Conclusion

We can conclude from our study that our proposed Hybrid ARIMA-FFT model is quite efficient in generating the short-term forecasts. Also, we have seen how our suggestion of modeling on the data starting from the first phase of lockdown, i.e. Datatype-2 as per the naming in our paper, generates more accurate forecasting models as far as the COVID-19 outbreak in India is concerned. The short-term forecasts will be helpful making future plans and taking few short-term decisions like whether to increase the lockdown or not, or how strict should the lockdown be depending on the trend.

References

- Box, G. E.P. (June 30, 2008): *Time Series Analysis: Forecasting and Control*. Wiley.
- COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). *ArcGIS*. Johns Hopkins University. (May 17, 2020).
- Frigo, M., and Johnson, S.G. A Modified Split-Radix FFT With Fewer Arithmetic Operations. *IEEE Transactions on Signal Processing*, **55(1)**, 111–119.
- Fumi, A., Pepe, A., Scarabotti, L. and Schiraldi, M.M. (2013): Fourier analysis for demand forecasting in fashion company. *International Journal of Engineering Business Management*. doi:10.5772/56839
- Heideman, M. T., Johnson, D.H. and Burrus, S. (1984): Gauss and the history of the fast Fourier transform. *IEEE ASSP Magazine*. doi:10.1109/MASSP.1984.1162257.
- Home | Ministry of Health and Family Welfare | GOI. *mohfw.gov.in* (17 May, 2020).
- Hunter, J.D. (2007): Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, 9, 90-95. doi:10.1109/MCSE.2007.55
- Hyndman, R.J. and Athanasopoulos G (2018): *Forecasting: Principles and Practice*. Otexts.
- India's Coronavirus Lockdown: What It Looks Like When India's 1.3 Billion People Stay Home. *Ndtv.com*. (Retrieved 11 April 2020).
- Kent, R.D., & Read, C. *Acoustic Analysis of Speech*. ISBN 0-7693-0112-6.
- McKinney, W. (2011): Pandas: a foundational Python library for data analysis and statistics, *Python for High Performance and Scientific Computing*.
- Miller, C. (2019): *Training Systems Using Python Statistical Modelling*. Packt Publishing Ltd.
- Pal, A. and Prakash, PKS. (September 28, 2017): *Practical Time Series Analysis: Master Time Series Data Processing, Visualization, and Modeling Using Python*. Packt Publishing.
- Seabold, S. and Perktold, J. “Statsmodels: Econometric and statistical modeling with python.” *Proceedings of the 9th Python in Science Conference*. 2010
- Strang, G. Wavelets. *American Scientist*, **82(3)**: 250–255. JSTOR 29775194.

Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). (January 30, 2020). *World Health Organization* (WHO).

Wolpert, D. H. and Macready, W. G. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, **1(1)**:67–82, 1997.

WHO Director-General's opening remarks at the media briefing on COVID-19- 11 March 2020. (Retrieved 11 March 2020). *World Health Organization*.

Authors and Affiliations

Mehuli Paul¹ and Meghanto Majumder²

Meghanto Majumder
meghanto20@gmail.com

¹M.Sc. Economics, Gokhale Institute of Politics and Economics, Pune, India-411004.

²B.Tech. Computer Science, St. Thomas' College of Engineering and Technology, Kolkata, India-700023.