<The End of the COVID-19 is Coming>

# Introduction

The end of COVID-19 guarantees tremendous opportunities in various fields, including businesses and employment; local restaurants could re-open, and temporarily laid-off airline employees could return to their workplace. According to Statista, about 413 million doses of COVID-19 vaccines had been produced up to March 2021, and China, U.S. EU, India, and U.K. are the top five countries that had produced 96% of all doses. However, these countries strictly control exportation of the vaccines for protecting their national economy. When the five countries no longer have new cases of COVID-19, the vaccines are expected to be commonly available around the world, and the pandemic is likely to end. This project applies Predictive Analytics and Time-Series Analysis, using auto.arima function in R, to predict and visualize a timeline for the end of the pandemic by determining when the COVID-19 trend stops in the five countries.

# Literature Review

F. Petropoulous, S. Makridakis, and N. Stylianou [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717777)] proposed a short-term forecasting model to predict two variables with regard to COVID-19, the cumulative number of confirmed cases and cumulative number of deaths. They followed a Rolling Origin Evaluation of Tashman (2000) to split the dataset into training sets and test sets for their model. The result graph contains the point forecasts, and the prediction intervals of 50, 70, and 90% for the confirmed cases. The authors emphasize that the uncertainty measures are as important as accuracy measures and claim that other papers do not provide uncertainty metrics in the literature review. The method they used to measure the uncertainty is to find the percentage difference between the predicted value and the 50% upper prediction interval, divided by the predicted value.

E. Gecili, A. Ziady, and R. Szczesniak [[2](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244173)] conducted four different time series models, the Holt model, the Autoregressive Integrated Moving Average (ARIMA) model, the TBATS, and the cubic smoothing spline model, to forecast the number of new COVID-19 cases for the USA and Italy. The study covers five periods of 7-day ahead point forecasts and prediction intervals, starting April 2, 2020. The prediction performance of all four models was similar; however, the ARIMA model and the cubic smoothing spline model were preferable because they had smaller predictions errors.

L. Ismail et al. [[3](https://www.sciencedirect.com/science/article/pii/S2001037020303998)] compared different time series forecasting models that are used to predict epidemics of infectious diseases. They applied and compared 17 different time series forecasting models to forecast the spread of COVID-19. They have developed a table that chooses an appropriate time series model to use, in terms of accuracy and uncertainty, according to its data trend. For example, in the countries where the dataset had an exponentially increasing curve, the ARIMA model outperformed the other models, yielding the least root-mean square error (RMSE) and the mean absolute percentage error (MAPE), while accurately forecasted the number of confirmed COVID-19 cases.

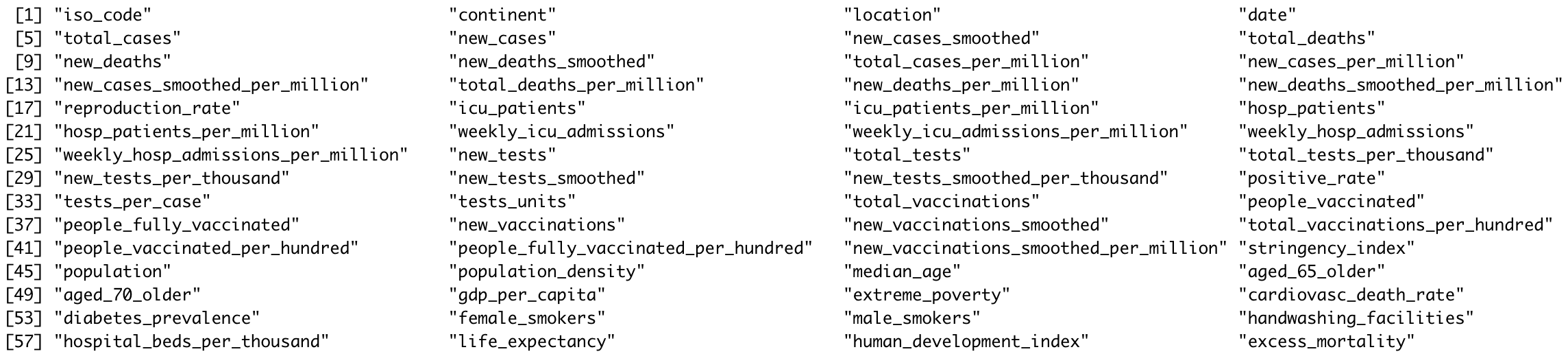
S. Shaikh et al. [[4](https://ieeexplore.ieee.org/document/9377137)] employed linear and polynomial regression to forecast the number of confirmed, active, cured, and death cases of COVID-19 in India, from March 12 to October 31, 2020. In the data processing, the data was split into 75% training set and 25% testing set, and given the attributes were the dependent variable, Y and “Months” was the independent variable, X. The accuracy and uncertainty were measured with R2 score, and MAPE, respectively. The study concluded that polynomial regression performed better than linear regression.

S. Maurya and S. Singh [[5](https://ieeexplore.ieee.org/document/9298390)] conducted and compared four different models, Naïve Model, Holt’s Linear Trend Method, Holt’s Winter Seasonal Model, and ARIMA Model, for their time series analysis of the COVID-19 datasets. This paper also explains when each model should be used in terms of the trend of the dataset. In their study, the Naïve model was the best model that had the least error value.

H. Elmousalami and A. Hassanien [[6](https://arxiv.org/abs/2003.07778)] developed time series forecasting models to predict global confirmed cases, global recovered cases, and global deaths cases of COVID-19. Single Exponential Smoothing (SES) had a far less RMSE and MAPE than a Moving Average model (MA) and Weighted Moving Average model (WMA) for forecasting confirmed, recovered and death cases.

# Dataset

The COVID-19 related data is publicly available from [Our World in Data](https://ourworldindata.org/covid-deaths). The data has 95,743 rows and 60 columns. Below is a list of columns of the data.



We will clean the data and extract the relevant variables that we need to conduct the time series analysis. The new data frame will consist of location (China, U.S., EU, India and U.K.), date, from 2020-02-01 to 2021-06-01, new cases, new deaths. Since our research question is to forecast new cases and deaths of COVID-19 in the five countries, the other variables will not be used.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Definition** |
| Location | Character | Country’s name |
| Date | Date | Date |
| New\_cases | Numerical | Number of confirmed cases of COVID-19 in a given date |
| New\_deaths | Numerical | Number of deaths cases of COVID-19 in a given date |

Table 1. Definition of variables

# Approach

## Step 1: <Pre-process Data in R>

Loading all the COVID-19 daily reports in R would be the first step; the reports are available in a single GitHub folder, written in csv. The rows that are not China, U.S. EU (Germany, Belgium, and Netherlands), India, and U.K., in each dataset, will be removed. Unnecessary columns will be removed and certain period of time will be selected for the study. Then, bind the datasets by country in ascending order.

## Step 2: <Exploratory Data Analysis>

Exploratory Data Analysis (EDA) is needed to gather insights from the dataset. First, we will check if there are gaps or missing values. In R, the str() function shows the total number of observations and its data type, and head() and tail() show the first and last five observations. The summary() function returns various summary statistics; it shows the minimum value, 1st quartile, median, mean, 3rd quartile and maximum value of the data. We can find how strong the correlation between “confirmed” and “deaths” of COVID-19. Patterns in the data must be studied. The decomposition check will allow us to study the trend, seasonality and irregular fluctuations of the dataset.

## Step 3: <Predictive Modeling>

In order to use an autoregressive integrated moving average, or ARIMA, model, the data must be stationary, univariate and in time series data format. If the data is not stationary (have a constant mean and variance), we will have to difference the data by subtracting the next value by the current value and make the data follow a stationary pattern. Autocorrelation function (ACF) and Partial autocorrelation function (PACF) will be used to determine the parameters, auto regression, integration and moving average, of an ARIMA model. Then we will use the auto.arima function with the three parameters to return a model and predict the number of new cases and deaths of COVID-19. We will conduct different n-day forecast; we can test our model, by comparing with actual values, using 30-day forecast; we can predict the future, using 90-day forecast.

## Step 4: <Performance Evaluation>

After a time series analysis, we can calculate the forecast error as the actual value minus the predicted value for each prediction. We can find the mean squared error (MSE) by taking the average of the squared forecast errors, then find the root mean squared error (RMSE) by square rooting the MSE. RMSE will measure the average magnitude of the error of the time series analysis. Another method to evaluate the performance of time series analysis is the mean absolute error (MAE), which can be calculated as the average of the absolute value of the forecast errors.

## Step 5: <Visualizing Results & Conclusion>

The ARIMA models display projection graphs forecasting new cases and deaths of COVID-19 for five countries. The main purpose of this project is to find if end of the COVID-19 is coming, with the help of vaccines. We can conclude that the pandemic will end soon if the future COVID-19 trend is downward and approaching towards zero.

# Results

In our time series analysis, auto ARIMA function was used to determine the best combination of ARIMA parameters in terms of AIC (Akaike Information Criterion) value. Auto ARIMA also transformed the non-stationary data to stationary data. Since all countries are not in the same condition, we analyzed various combinations of the frequency of seasonality and the length of training set for each country. Four different models, that predict the number of new cases of COVID-19 after one month, were built to compare and select the best model to predict when new case number reaches to zero. In order to evaluate their measures, distance (difference), mean absolute percentage error (MAPE) and root mean squared error (RMSE) were compared.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Actual Value | Predicted Value | Difference | % in Difference | MAPE | RMSE | Selected model |
| US (training set of 2020/03 – 2021/06), frequency = 7 | 14,463 | 8,893.08 | 5,569.91 | 39% | 42.17 | 5,263.46 |  |
| US (training set of 2021/01 – 2021/06), frequency = 7 | 14,463 | 4,236.65 | 10,226.35 | 71% | 42.90 | 5,321.23 |  |
| US (training set of 2020/03 – 2021/06), frequency = 12 | 14,463 | 19,618.77 | 5,155.77 | 36% | 35.65 | 5,155.77 |  |
| US (training set of 2021/01 – 2021/06), frequency = 12 | 14,463 | 19,318.81 | 4,855.81 | 34% | 33.57 | 4,855.80 | ✓ |

Table 2. Comparison of times series models for US

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Actual Value | Predicted Value | Difference | % in Difference | MAPE | RMSE | Selected model |
| China (training set of 2020/03 – 2021/06), frequency = 7 | 18 | 29.76 | 11.76 | 65% | 75.13 | 10.94 |  |
| China (training set of 2021/01 – 2021/06), frequency = 7 | 18 | 29.7 | 11.70 | 65% | 69.78 | 10.85 |  |
| China (training set of 2020/03 – 2021/06), frequency = 12 | 18 | 23.33 | 5.33 | 30% | 29.59 | 5.3263 |  |
| China (training set of 2021/01 – 2021/06), frequency = 12 | 18 | 19.86 | 1.86 | 10% | 10.35 | 1.86 | ✓ |

Table 3. Comparison of times series models for China

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Actual Value | Predicted Value | Difference | % in Difference | MAPE | RMSE | Selected model |
| India (training set of 2020/03 – 2021/06), frequency = 7 | 46,617 | 98,886.52 | 52,269.52 | 112% | 91.39 | 52,880.40 |  |
| India (training set of 2021/01 – 2021/06), frequency = 7 | 46,617 | 91,799.52 | 45,182.52 | 97% | 82.69 | 48,218.16 | ✓ |
| India (training set of 2020/03 – 2021/06), frequency = 12 | 46,617 | 143,780 | 97,163 | 208% | 208.43 | 97,163.18 |  |
| India (training set of 2021/01 – 2021/06), frequency = 12 | 46,617 | 118,705 | 72,088 | 155% | 154.64 | 72,088.03 |  |

Table 4. Comparison of times series models for India

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Actual Value | Predicted Value | Difference | % in Difference | MAPE | RMSE | Selected model |
| UK (training set of 2020/03 – 2021/06), frequency = 7 | 28,071 | 3,944.24 | 24,126.76 | 86% | 63.49 | 13,806.76 |  |
| UK (training set of 2021/01 – 2021/06), frequency = 7 | 28,071 | 4,175.98 | 23,895.02 | 85% | 61.48 | 13,663.83 | ✓ |
| UK (training set of 2020/03 – 2021/06), frequency = 12 | 28,071 | 4,599.55 | 23,471.45 | 84% | 83.61 | 23,471.45 |  |
| UK (training set of 2021/01 – 2021/06), frequency = 12 | 28,071 | 4,283.79 | 23,787.21 | 85% | 84.74 | 23,787.21 |  |

Table 5. Comparison of times series models for UK

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Actual Value | Predicted Value | Difference | % in Difference | MAPE | RMSE | Selected model |
| EU (training set of 2020/03 – 2021/06), frequency = 7 | 2,337 | 7,650.10 | 5,313.10 | 227% | 293.51 | 7,342.32 |  |
| EU (training set of 2021/01 – 2021/06), frequency = 7 | 2,337 | 5,705.73 | 3,367.73 | 144% | 254.48 | 6,719.08 | ✓ |
| EU (training set of 2020/03 – 2021/06), frequency = 12 | 2,337 | 10,339.34 | 8,002.34 | 342% | 342.42 | 8,002.34 |  |
| EU (training set of 2021/01 – 2021/06), frequency = 12 | 2,337 | 12,023.28 | 9,686.28 | 414% | 414.47 | 9,686.28 |  |

Table 6. Comparison of times series models for EU

For US, China, India and EU, the model with the closest distance has the least MAPE and RMSE. On the other hand, the UK’s second model has the least MAPE and RMSE, but the distance is not the closest. The second model was selected because there is no significant difference in the distance, but it has 26% lower MAPE and 42% lower RMSE than the third model. It is observed that all models have a better result when the training set is five months rather than 15 months.

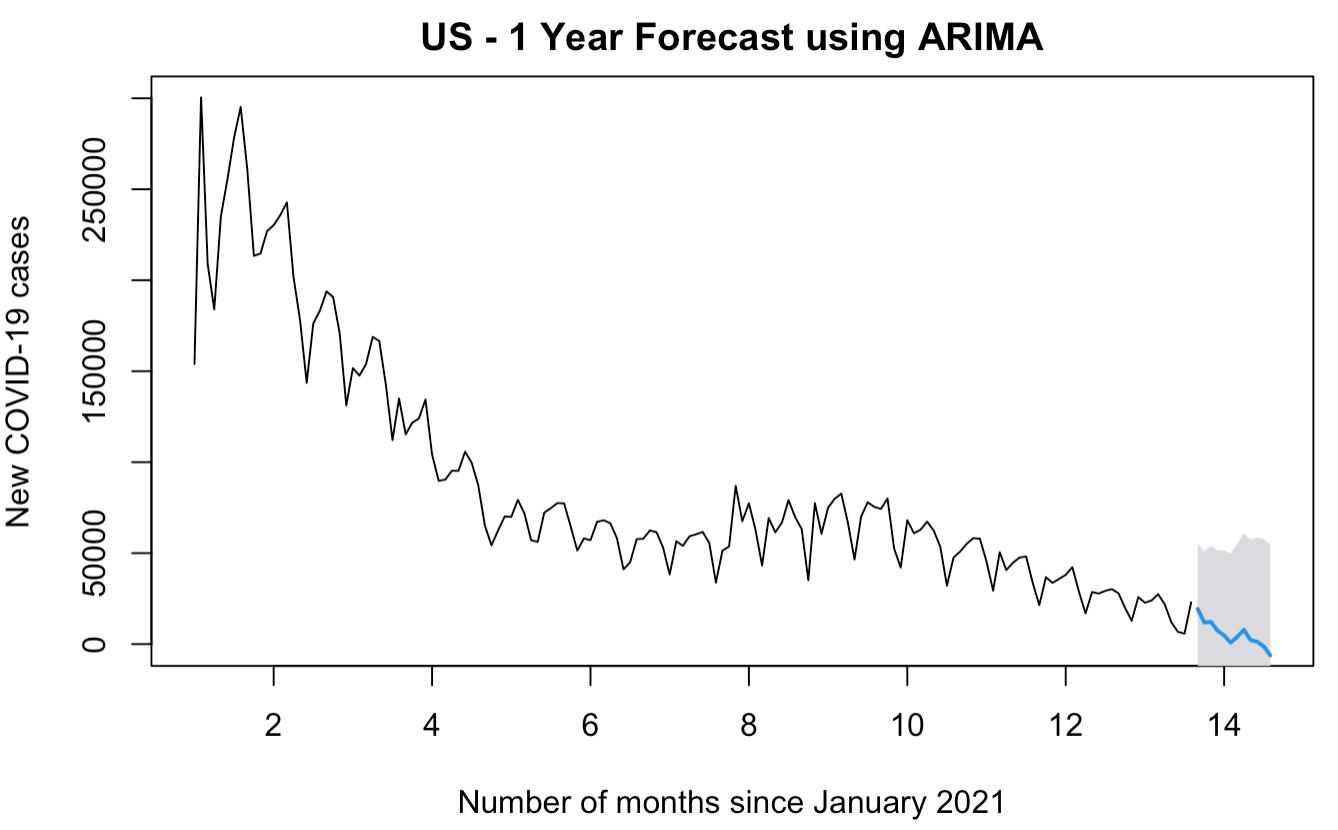
After the COVID-19 vaccines have been produced in March 2021, all countries, except for China, have a decreasing COVID-19 trend. Based on the past five months of data, our models forecast that the spread of COVID-19 in US, India, UK and EU stops in the near future.

Fig 1. Plot graph of US ARIMA model

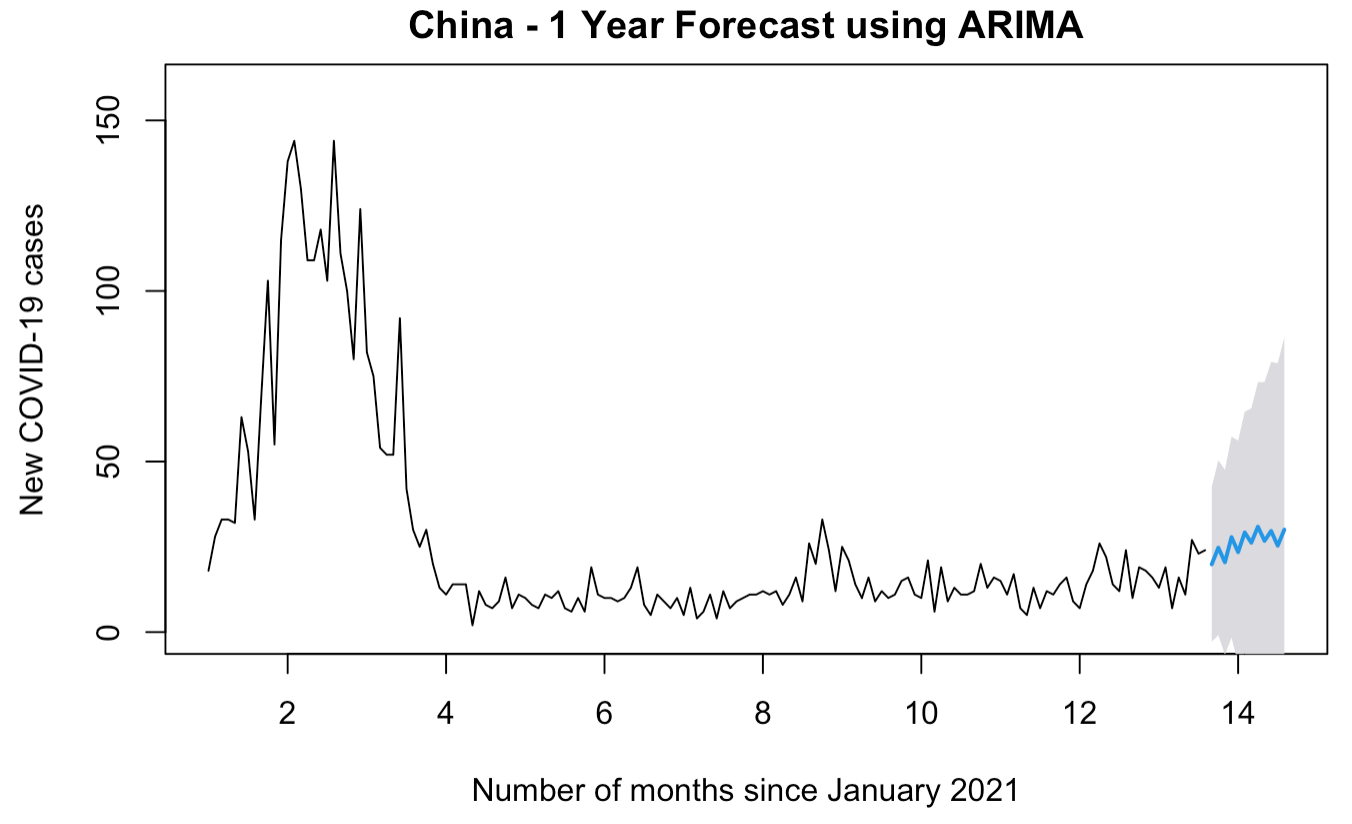
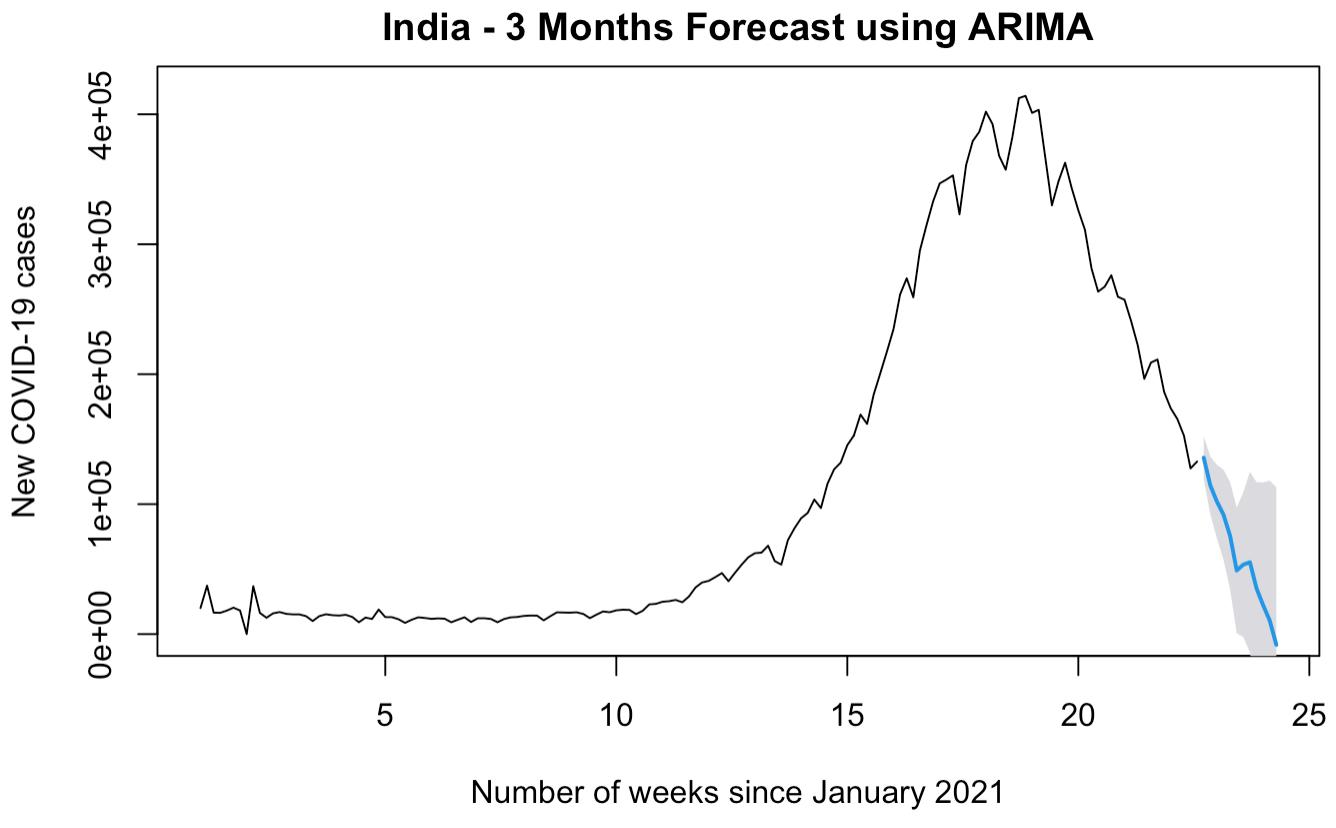


Fig 2. Plot graph of China ARIMA model

Fig 3. Plot graph of India ARIMA model

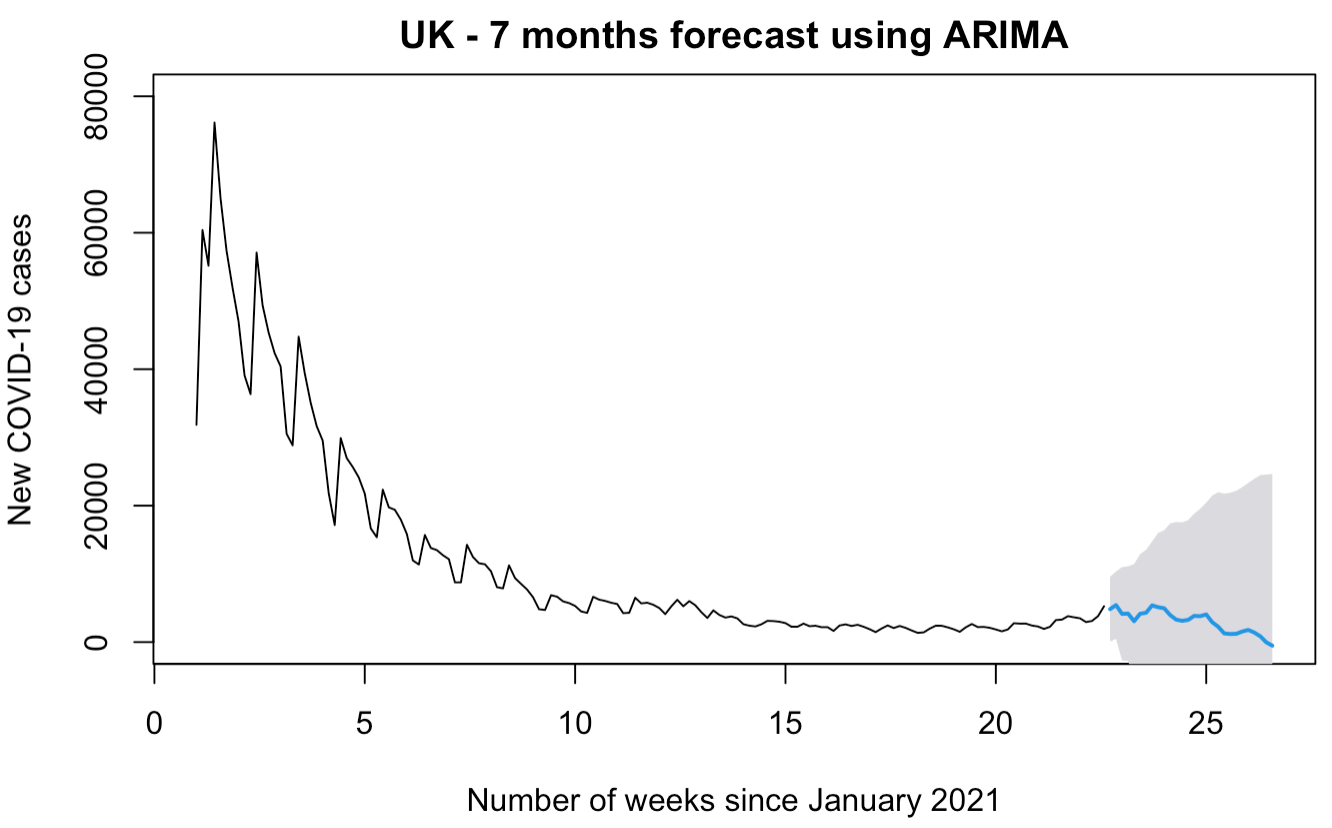
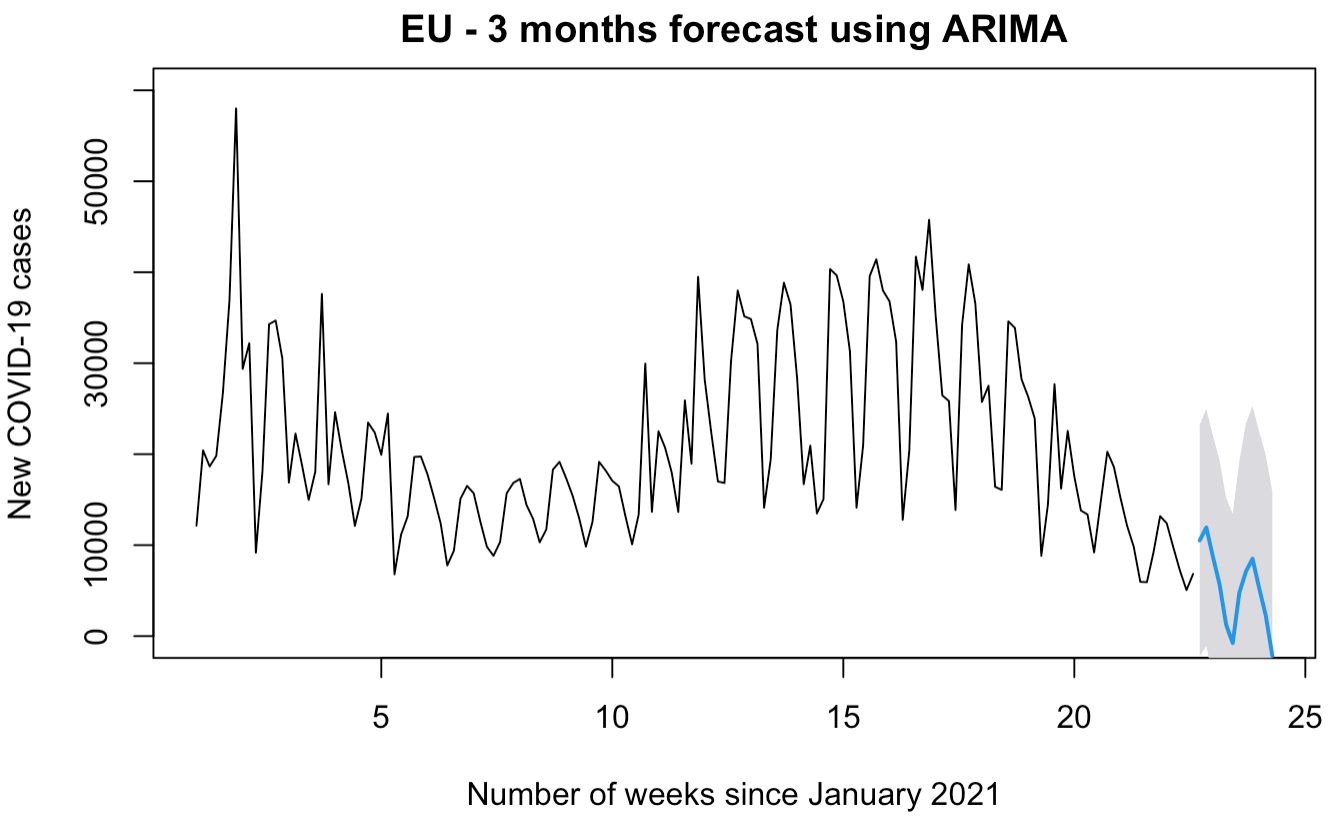


Fig 4. Plot graph of UK ARIMA model

Fig 5. Plot graph of EU ARIMA model

The predicted number of new COVID-19 cases for each country is provided below table. Our time series models predict that the COVID-19 trend in US, India, UK and EU will end in one year, three months, six months, and three months, respectively. The recent numbers of new cases in China was increasing, so the model does not forecast that the new case in China reaches zero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted Value for August 1 | Predicted Value for September 1 | Predicted Value for October 1 | Estimated End Date |
| US | 11899 | 12195 | 7370 | May 2022 |
| China | 25 | 20 | 28 | unknown |
| India | 55310 | 0 | 0 | September 2021 |
| UK | 5381 | 3295 | 3800 | January 2022 |
| EU | 7096 | 0 | 0 | September 2021 |

Table 7. Forecast of next three months and estimated end date

## Conclusions

In the meanwhile, new COVID-19 virus variants are developed in some countries, and the vaccines do not work against some variants. Forecasting the pandemic is a challenging project as indefinite external factors influence its infectiousness. In the past three months, we proposed live COVID-19 related predictions for five vaccine producing countries. Our proposed time series models are selected, among various combinations, based on good levels of performance and uncertainty. The results should be more accurate if the COVID-19 trend was the same or similar as three months ago.

The pandemic may not end this year or next year as our models forecasted, due to the Delta variant. In order to improve, our models should keep track of new COIVD-19 trend and accumulate more recent data. It will eventually forecast the end of the pandemic, again, when new vaccines come out.

## References

1. Petropoulos, F., Makridakis, S., & Stylianou, N. (2020). COVID-19: Forecasting confirmed cases and deaths with a simple time series model. *International Journal of Forecasting*. Published. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717777/>
2. Gecili, E., Ziady, A., & Szczesniak, R. D. (2021). Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLOS ONE*, *16*(1), e0244173. https://doi.org/10.1371/journal.pone.0244173
3. Ismail, L., Materwala, H., Znati, T., Turaev, S., & Khan, M. A. (2020). Tailoring time series models for forecasting coronavirus spread: Case studies of 187 countries. *Computational and Structural Biotechnology Journal*, *18*, 2972–3206. https://www.sciencedirect.com/science/article/pii/S2001037020303998
4. Shaikh, S., Gala, J., Jain, A., Advani, S., Jaidhara, S., & Edinburgh, M. R. (2021). Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. doi:10.1109/confluence51648.2021.9377137
5. Maurya, S., & Singh, S. (2020). Time Series Analysis of the Covid-19 Datasets. *2020 IEEE International Conference for Innovation in Technology (INOCON)*. doi:10.1109/inocon50539.2020.9298390
6. Elmousalami, H. H., & Hassanien, A. E. (2020, March 15). Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modeling and Recommendations. Retrieved from https://arxiv.org/abs/2003.07778