https://doi.org/10.29013/YSJ-22-5.6-19-25

Joanna Yao,
General Studies,
Millburn High School, Millburn, New Jersey, U.S.A.

A COVID-19 DATA ANALYSIS IN A STATIONARY TIME SERIES

Abstract. This paper starts with a general discussion regarding the spread of COVID-19 and aims to use R to make regional predictions regarding the virus's spread. We carried out multiple lines of code of time series models using data sets from different countries, specifically New Jersey, United States, and Shanghai, China. We sought to analyze patterns in the virus's proliferation to make future forecasts. The results demonstrate that the model with the coefficients $(1, 0, 0)$ can be useful to forecast the number of COVID-19 cases. This statistical forecast can be helpful for current and future resources allocation and epidemic prevention, as well as epidemiology and disease study.

Keywords: COVID-19; epidemiology; prediction; forecasting; ARIMA model; time series; statistics; public health.

I. Introduction

1.1 Background

The coronavirus SARS-CoV-2 proliferated globally; as of September 2021, countries such as the United States, India, and the United Kingdom have been impacted severely, most noticeably densely-populated cities. Through respiratory droplets from sneezing and coughing, the virus typically incubates for five to seven days at a maximum of fourteen days and causes initial symptoms such as fever, cough, nasal congestion, and fatigue. More symptoms following infection progression include severe chest symptoms and viral pneumonia, accompanied by decreased oxygen saturation, lymphopenia, and elevated inflammatory markers.

COVID-19 was first detected in Wuhan, China in late 2019 [3]. As of March 1, 2020, 79968 patients in China and 7169 patients outside of China have been diagnosed with the virus, and as of November of 2021, 248 million people globally have had COVID-19, with a little over 5 million deaths [2]. Medical professionals consider patients older than sixty years old at higher risk [5] and in February 2020, estimated an average fatality rate of approximately 2.2% [7], which depends on factors such as age and immunity. Aside from posing as an international public health danger, COVID-19 has also impacted other aspects of society, including education, legal proceedings, and work life — this research itself has been conducted virtually due to the pandemic. The pandemic has been an offset to the sense of normalcy; hence, this research's objective is to determine data that may help community health services to keep the pandemic's spread to a minimum.

1.2 Research Objectives

This research analyzes the number of COVID-19 cases for various countries and its progression based on the number of forecasted cases in 2020. By focusing specifically on New Jersey and Shanghai, mathematical analysis can potentially develop models, which can forecast the number of cases in a certain amount of days. With such tools, government officials would be able to estimate the severity of COVID-19 in the future and implement necessary protective procedures or medical supply distributions to maintain the spread of the virus to a minimum and assist marginalized and underprivileged communities, as well as analyze the progression of the virus's proliferation visually.

II. Methodology

2.1 Dataset Overview

The training dataset consists of six columns: Forecast ID, province/state, country region, date of record, confirmed cases, and fatalities. Each row lists a unique identifier, the specific region of the country (if applicable), the country itself (which is organized in alphabetical order), and the number of confirmed cases and fatalities on that day. As time passes, the number of COVID-19 cases clearly increases with time. This research applies the training dataset to creating and training an initial model for forecasting COVID-19 cases.

The validation dataset records data for four columns: Forecast ID, province/state, country region, and date of record, each row with a unique identifier for every day, correlating with the data from the training dataset. Using the validation set fine-tunes the model to handle future data. From what we can see, as time passes, the number of forecasted cases increases.

The testing dataset displays three columns recording the number of forecasted COVID-19 cases, the number of confirmed cases, and the number of fatalities. By running this test data through the model, we can see how accurate the predicted outputs are.

Shown is the data visualization plot of the training set comparing the increase of confirmed cases and fatalities globally. The number of confirmed cases increased much faster than the number of fatalities. Particularly, after mid-March of 2020, the number of confirmed cases spiked rapidly.

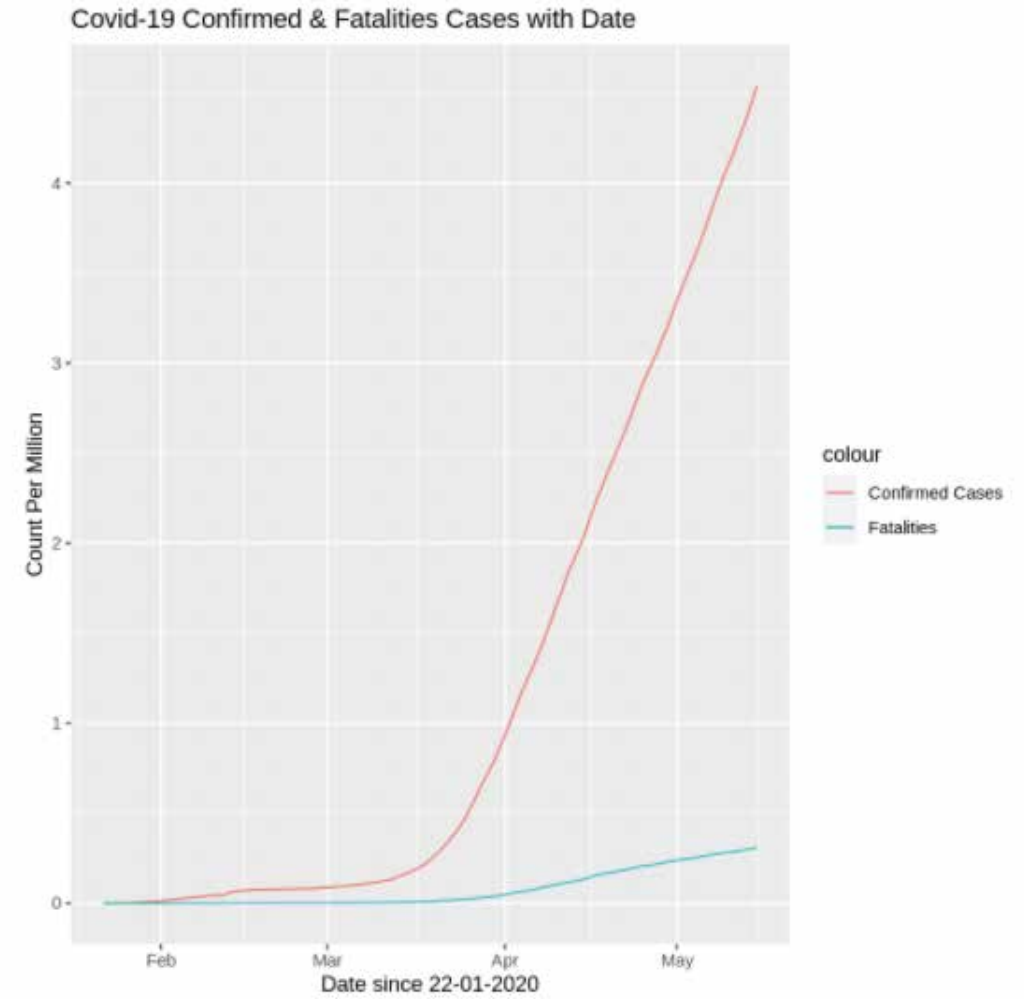

Figure 1. Count of COVID-19 cases

Below, dotted visualization of the progression of confirmed cases in several different countries from the training set is presented. As shown with the steep-sloped pink line representing the U.S, there is a far more significant increase in confirmed COVID-19 cases. Earlier in 2020, in February, the number of confirmed cases in China seemed to increase quite a bit before leveling off for the duration of March to May. Towards the latter half of March, however, the United States especially experienced a far more severe increase in confirmed cases, correlating with the visual above. The number of confirmed cases in countries including Italy, Iran, South Korea, and Germany also appear to increase sharply, though less than U.S., around the end of March, which could explain how the number of confirmed cases globally proliferated around that period of time.

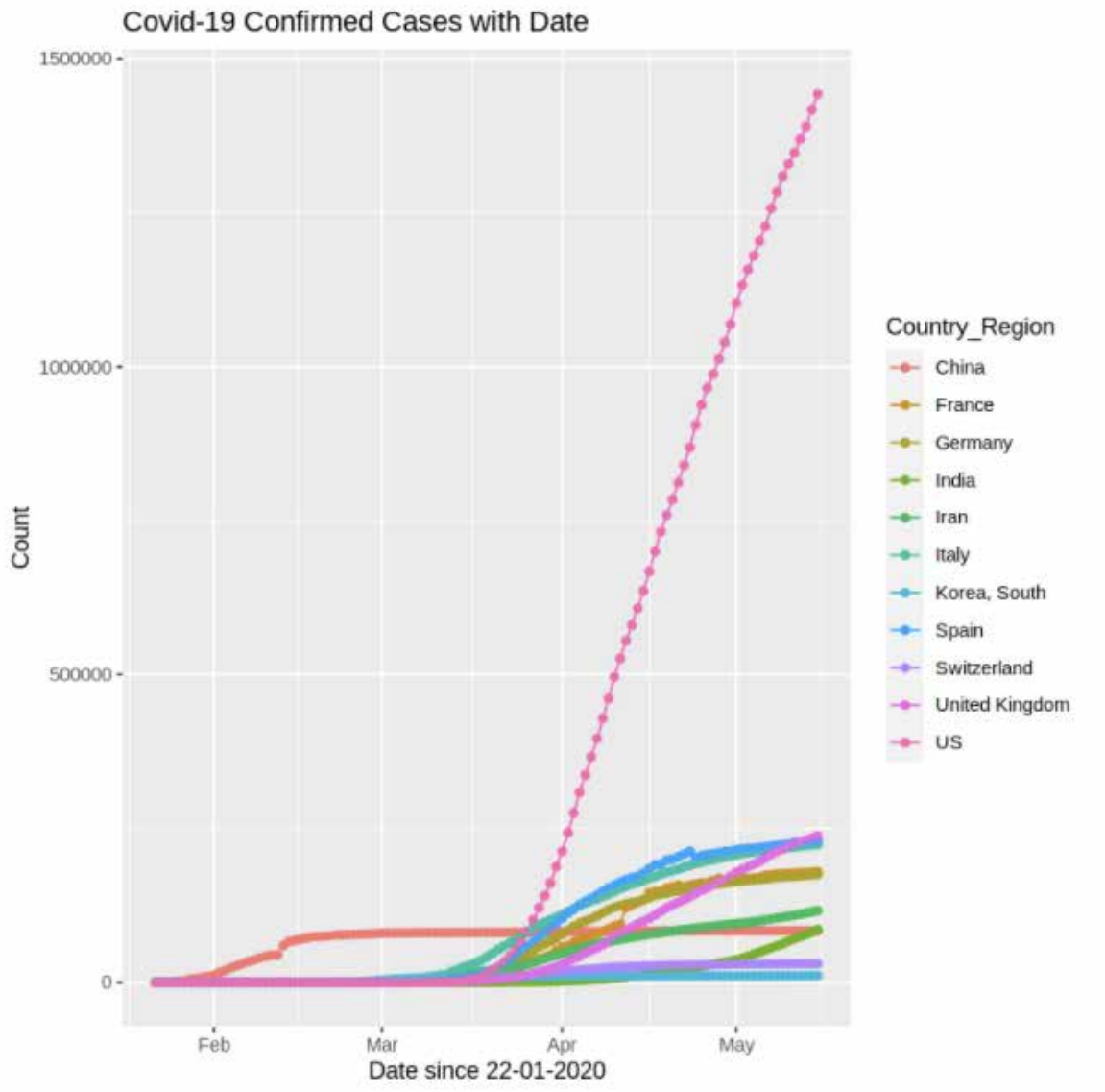

Figure 2. COVID-19 progression by region

The last visualization exhibited is a bar graph comparing the number of confirmed cases and fatalities in several countries. Both the number of confirmed cases and the number of fatalities in the U.S. were significantly higher than those of other countries.

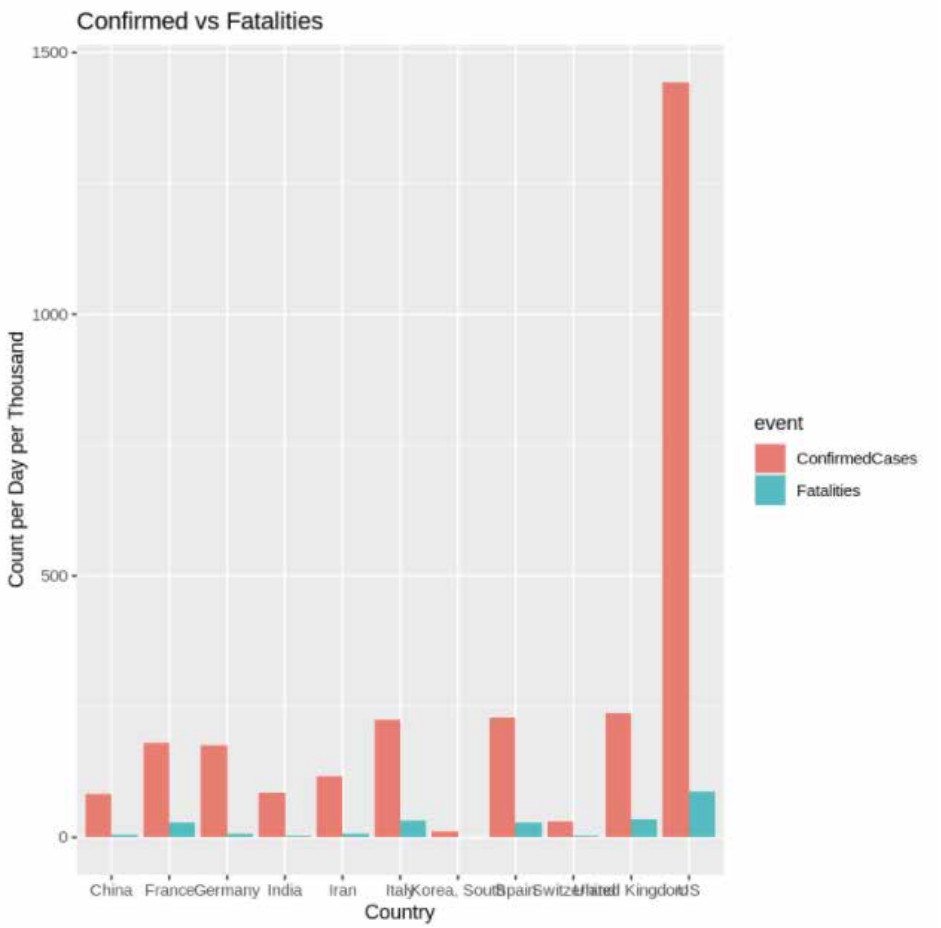

Figure 3. Comparison of conirmed coses and fatalities by region

2.2 The Importance of a Stationary Series

For this paper, a normal time series is not sufficient since if used, the model and its variance would be a function of time. As time passes, each nation's COVID-19 case count will increase, but the ratio of confirmed cases and death cases varies a lot among different countries. Although a visualization of this information could display trends and the degree of severity of disease proliferation, its reliability in predicting the number of future cases is subpar.

This research will analyze a stationary time series, which are what most forecasting methods are designed for, and measure the autocorrelation — the constant relationship between a variable's current value and past values.

2.3 Time Series Models and Predictions

This research will construct ARIMA models (autoregressive integrated moving average) to forecast the number of COVID-19 cases, which can be made stationary by differencing. An ARMA model is a combination of the aspects of AR (autoregressive) models, which utilize previous data values to predict future values, and MA (moving average) models, which remains constantly stationary, and alongside integration [4]. This type of model is most suitable since it is a general class of forecasting models that can be transformed to be stationary through differencing, logging, or deflating.

Focusing on New Jersey, United States, and Shanghai, China, the research aims to predict the number of COVID-19 cases in those regions in the future. By focusing on regions in two different continents, the research also hopes to investigate the inevitable effects of different population densities on the local proliferation of COVID-19.

III. Implementation

3.1 Model Construction

Via R Studio, two separate models were generated for Shanghai, China and New Jersey, United States. The datasets were first cleaned up to ensure each row had a "providence" name just for consistency. Specifically, here is what was yielded for the first few rows of the data after cleaning it up:

```
> head(covid_training_dataset)
  Id Province_State Country_Region       Date ConfirmedCases Fatalities
1  1    Afghanistan    Afghanistan 2020-01-22              0          0
2  2    Afghanistan    Afghanistan 2020-01-23              0          0
3  3    Afghanistan    Afghanistan 2020-01-24              0          0
4  4    Afghanistan    Afghanistan 2020-01-25              0          0
5  5    Afghanistan    Afghanistan 2020-01-26              0          0
6  6    Afghanistan    Afghanistan 2020-01-27              0          0
> tail(covid_training_dataset)
          Id Province_State Country_Region       Date ConfirmedCases Fatalities
35990 35990       Zimbabwe       Zimbabwe 2020-05-10             36          4
35991 35991       Zimbabwe       Zimbabwe 2020-05-11             36          4
35992 35992       Zimbabwe       Zimbabwe 2020-05-12             36          4
35993 35993       Zimbabwe       Zimbabwe 2020-05-13             37          4
35994 35994       Zimbabwe       Zimbabwe 2020-05-14             37          4
35995 35995       Zimbabwe       Zimbabwe 2020-05-15             42          4
```

Figure 4. Displayed below are the first and last six rows of the training dataset with the "Province_State" heading filled in, as in the original dataset, it was missing

Afterwards, the model to forecast COVID-19 cases in Shanghai was created with the Arima() function of the training dataset to forecast the number of confirmed COVID-19 cases in 43 days, which was rooted from April 2nd, 2020. Specifically, an ARIMA model with the coefficients of (1, 0, 0) were used to formulate the predictions because it is considered to be "first-order", and due to the consistently increasing nature of the COVID-19 cases, it could be assumed that a multiple of the previous entry alongside an additional constant would suffice as a mathematically-based prediction.

Afterwards, with the print() function, the program printed out the results of the formula in a table similar to that of the dataset's format, except the output produced the predictions that resulted from the ARIMA model. Below, the first view rows of these predictions are shown:

```
  ForecastId Province_State Country_Region       Date ConfirmedCases_predict ConfirmedCases Fatalities
1       3398       Shanghai          China 2020-04-02                     3              1          1
2       3399       Shanghai          China 2020-04-03                     1              1          1
3       3400       Shanghai          China 2020-04-04                     1              1          1
4       3401       Shanghai          China 2020-04-05                     1              1          1
5       3402       Shanghai          China 2020-04-06                     0              1          1
6       3403       Shanghai          China 2020-04-07                     2              1          1
  ForecastId Province_State Country_Region       Date ConfirmedCases_predict ConfirmedCases Fatalities
1      11525     New Jersey             US 2020-04-02                   707              1          1
2      11526     New Jersey             US 2020-04-03                  1592              1          1
3      11527     New Jersey             US 2020-04-04                   268              1          1
4      11528     New Jersey             US 2020-04-05                  1399              1          1
5      11529     New Jersey             US 2020-04-06                  1542              1          1
6      11530     New Jersey             US 2020-04-07                   806              1          1
```

Figure 5. The first 6 rows of the predictions generated by the ARIMA model for Shanghai, China, and New Jersey, United States (respectively) starting April 2nd, 2020. This was generated by the head(result) function

IV. Results

4.1 Model Limitations:

While these models demonstrate reasonable accuracy, they are restricted. The model parameters are targeted for a specific region or country. This research only designed them for two regions in two different countries, which is rather small-scaled compared to the rest of the world. The data produced would not apply for other countries and therefore would not be useful for government officials of other nations to use to enforce disease control or regulations in their own regions. Factors that vary between regions such as population density of certain cities and the number of vaccines available per country are why a model based on data from one country cannot apply to another.

Secondly, the research constructed the models based on data from January 2020. The models are not fully up-to-date, and since the pandemic is not yet over, there is not much data processed for more recent days. The number of COVID-19 cases would become unpredictable without the most recent data, but more time-series construction can be done. The COVID-19 pandemic is also quite similar to the SARS (severe acute respiratory syndrome) pandemic in terms of origin, having rooted from a coronavirus, and the statistics between these two pandemics could be compared, though it is noted that these two pandemics occurred at different times and that scientific research, technology, and communications has advanced over the decade, accounting for the differences in the mortality rates and spreads of these diseases. Especially with the rising of the Omicron variant, which is more transmissible, data from early 2020 fails to take that increase in proliferation into consideration, deeming the models as less up-to-date.

In the future, there ought to be more region-specific analyses throughout the world not limited to the two regions below, in which health officials could also utilize other forms of data analysis, potentially opening up a forecasting database.

4.2 Potential Model Applications

This research shows that ARIMA models can generate useful data about the potential COVID-19 case count and the fatalities. With a method of generating predictable data, governments can potentially collaborate with statistic and epidemiological professionals internationally to generate large databases providing information about the future of the virus's proliferation in certain regions. Healthcare officials can also use the calculated predictions as means to determine which nations have the fastest COVID-19 proliferation rates or fatality rates, keeping international organizations such as World Health Organization (WHO) and the REd Cross informed and possibly ready to take action if needed. Furthermore, if governments choose to make these databases public, citizens can remain well-informed regarding control protocol and safety and can make necessary arrangements.

The predictions estimated by the models are also useful to estimate the amount of vaccines/medical supplies needed for certain locations. For instance, in Burundi, the minister of health had approved the first order of COVID-19 vaccines to be distributed in late July following a rise in the spread of infectious variants [1]. Data such as the predicted number of COVID-19 cases generated by the model can forecast the virus's proliferation and can even be applied to new variants, such as the Omicron variant. This variant has already spread to more than six U.S. states, and such models can be applied in different locations so that authorities can identify areas that are most threatened and implement any necessary measures.

V. Conclusion

The number of COVID-19 cases per region may vary since China and New Jersey, two different regions, show different numbers of COVID-19 cases. Since that is the case, it may be harder to predict the number of cases in all of the regions holistically. Hence, creating two separate forecasts for two separate regions would provide authorities in each of those areas with more precise information.

While the models are reasonably accurate and can be used for data collection, revisions are imperative. As the Omicron variant becomes more prevalent and is said to be more easily transmissible, new data should be added to create new models that can predict the number of confirmed cases relative to the proliferation rate at the time. Since many universities and other public institutions are already going remote, it is a crucial time for this model to be considered as a means to plan ahead for another potential period of quarantine.

References:

1. Associated Press. "Burundi, in Reversal, Says It Will Accept COVID-19 Vaccines". U. S. News, last modified July 29, 2021. Accessed December 25, 2021. URL: https://www.usnews.com/news/world/articles/2021–07–29/burundi-in-reversal-says-it-will-accept-covid-19-vaccines.
2. David Baud et al. "Real Estimates of Mortality following COVID-19 Infection". The Lancet Infectious Diseases 20,— No. 7. 2020.: [Page #]. URL: https://doi.org/10.1016/S1473–3099(20)30195-X
3. "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". Infographic. Johns Hopkins University & Medicine. 2021. https://coronavirus.jhu.edu/map.html.
4. Robert Nau. "Lecture Notes on Forecasting". Lecture, last modified 2014. Accessed: October 8, 2021. URL: https://people.duke.edu/~rnau/Slides_on_ARIMA_models — Robert_Nau.pdf.
5. Graziano Onder, Giovanni Rezza and Silvio Brusaferro. "Case-fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy". The Journal of the American Medical Association 323,— No. 18. 2020. [Page #]. URL: https://doi.org/10.1001/jama.2020.4683
6. Varotsos Costas A. and Vladimir F. Krapivin. "A New Model for the Spread of COVID-19 and the Improvement of Safety". ScienceDirect. Last modified 2020. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925753520303593#.
7. Velevan Thirumalaisamy P. and Christian G. Meyer. "The COVID-19 Epidemic". National Center for Biotechnology Information. Last modified February 16, 2020. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7169770/.
8. Code for Replication, derived from R-Studio: URL: https://colab.research.google.com/drive/1g8YxUjOVnLzGFYHb5zrr2X7Y94Vrbd0u?usp=sharing