COVID-19 DATA ANALYTICS AND FORECASTING

Hang Le Thi Thuy[1], Vinh Nguyen Do Phu[2], Dai Nguyen Quoc[3], Nghiep Nguyen Le Tan[4] and An Bui Pham Thai[5]

Abstract

This project gives an overview of the most recent pandemic of the novel coronavirus 2019 (COVID-19) and shows how data analysis of such an epidemic plays an important role for further predictions and even predictions of future situations.

Key words: COVID-19, data analysis, forecasting model

1 Introduction

The key to understanding the pandemic begins with understanding the disease itself and how the disease progresses naturally. This report presents different categories of diseases and different routes of disease transmission. Various past pandemics and stages of the pandemics are also discussed.

Accurately predicting the rate of spread and infection can help minimize the outbreak by taking precautionary measures. However, forecasting requires data and there are various data processing challenges. The main objective of this project is to present and discuss a series of predictions about the behavior of the pandemic of the following three groups using data from July 1, 2020 to May 31, 2021:

(i) Global

(ii) Three countries with highest confirmed cases: United States, Brazil, India

(iii) Three Asian countries: Japan, Korea, Vietnam

[1] Faculty of Computer Science and Engineering, University of Technology, Vietnam National University, Ho Chi Minh city. hang.le2911@hcmut.edu.vn

[2] Faculty of Computer Science and Engineering, University of Technology, Vietnam National University, Ho Chi Minh city. vinh.nguyen16302@hcmut.edu.vn

[3] Faculty of Computer Science and Engineering, University of Technology, Vietnam National University, Ho Chi Minh city. dai.nguyenkhmtclc@hcmut.edu.vn

[4] Faculty of Computer Science and Engineering, University of Technology, Vietnam National University, Ho Chi Minh city. nghiep.nguyen0908918318@hcmut.edu.vn

[5] Faculty of Computer Science and Engineering, University of Technology, Vietnam National University, Ho Chi Minh city. an.bui1104@hcmut.edu.vn

One deterministic model widely considered in epidemiology is the Susceptible–Infectious–Removed (SIR) model, which is based on the classification of the individuals into three stages of infection and was introduced almost one hundred years ago by W.O.K. McKendrick and A.G. McKendrick. The transition rates from one class to another are mathematically expressed as derivatives, hence the model is formulated using differential equations. While building such models, it must be assumed that the population size in a compartment is differentiable with respect to time and that the epidemic process is deterministic. In other words, the changes in the population of a compartment can be calculated using only the history that was used to develop the model [1]. More details about the model are presented in the following sections.

Random Forest Regression (RFR) is a type of machine learning that can analyze complex interactions between clinical characteristics and provide high classification accuracy using a set of decision trees. [2]

Finally, ARIMA model will consider an intervention factor to make it possible to reflect external measures into the infection rates. [3]

These models must be considered as a tool to support the decision-making process, but may not reflect the effective rates of infection in medium and long-term analysis. Real-life models are subject to many other external and internal parameters. In this report, positive aspects and limitations of each model are presented.

2 Epidemiology terms and concept

In order to understand a pandemic / epidemic, one must first understand the disease itself and the course of the natural course of the disease.

2.1 Introduction

The word “disease” is defined as the condition that negatively affects the body of a living person, plant or animal. A disease affects a body due to a pathogenic infection. The natural course of the disease begins before the onset of infection and then passes through the pre-symptomatic stage. The final phase is the clinical phase. In the clinical phase, a patient receives the prognosis of the disease. After successful treatment of the disease, the patient enters into the remission stage. Remission refers to a decrease in the symptoms or a complete disappearance of the disease. The patient must strictly adhere to the doctor’s instructions during the remission phase. This will ensure that the disease does not recur. If treatment is unsuccessful, the patient may die or be chronically disabled. [4]
Diseases are mainly categorized into two types:

(i) *Congenital diseases*

Exist in the body right from birth. These diseases are generally activated through genetic disorders, environmental factors, or a combination of both. As a result, congenital disease are hereditary, i.e. passed on through generations, for example, hearing disorders and Down syndrome.

(ii) Acquired diseases

In contrast to the former, acquired diseases spread through living organisms. These are not hereditary.

The ***acquired disease*** category is further classified into two types:

(i) Infectious diseases

Infectious diseases are triggered by pathogens or viruses. They are also called *communicable diseases*. As the name suggests, these diseases are contagious. This means that a contagious disease in one person can be passed on to another person through air, food, water, touch (physical contact), etc. SARS and SARS COVID-19 are examples of infectious diseases.

(ii) Non-infectious diseases

Non-infectious diseases do not occur due to any type of infection. This means that a person with a non-infectious disease will not be able to spread the disease to a healthy person. Diseases such as cancer and autoimmune disorders are examples of non-infectious diseases.

Infectious disease can affect a healthy person in two ways:

(i) Direct transmission

When the pathogens travel from a patient to a healthy person without an intermediate carrier, this is referred to as direct transmission. Direct transmission can happen in the following ways:

- Coming in contact with the infected person.
- Via droplet infection (coughing, sneezing, and spitting).
- Coming in contact with the soil.

Animal bites are also one of the causes of direct transmission.

(ii) Indirect transmission

Whenever there is a reservoir of infection that can transmit the disease from a patient to a healthy person with a medium pathogen, this transmission is known as indirect transmission. Indirect transmission can happen in the following ways.

- If pathogens are transmitted through food, water, etc., it is known as vehicle-borne disease.
- If pathogens are transmitted through the air, then it is known as airborne disease.
- If pathogens are transmitted through contaminated items like clothing, utensils, books, etc., it is known as fomite-borne disease.

After the diagnosis of the disease comes the most important part: the treatment. Treatment generally consists of targeting the biochemical reactions occurring due to pathogens. There are two ways to stop that reaction so that the infection will not spread:

(i) Prevention

Through prevention, symptoms of the infection can be relieved using painkillers so that patients can be at ease. Preventive measures also include immunization and vaccinations.

(ii) Cure

Through cure, particular drugs are used to kill the pathogen. [5]

2.2 Overview of epidemic

2.2.1 Stages of Disease

Before studying the latest pandemic, it is very important to study basic terminologies associated with the pattern of disease spread. A diagrammatical overview of stages of the disease is depicted below [6]:

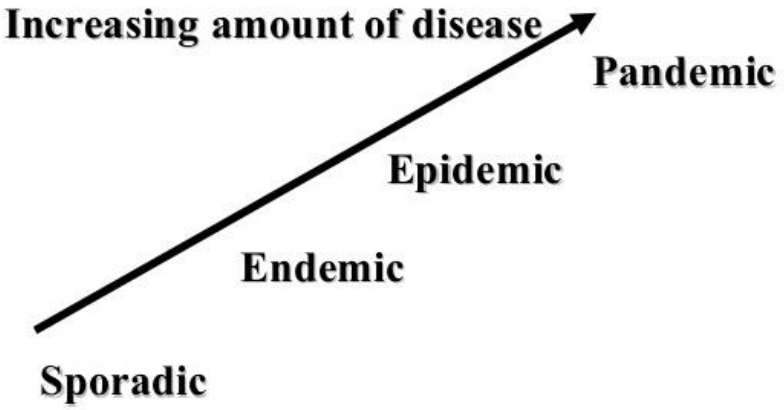


FIGURE 1. Progressive stages of a disease.

(i) Sporadic

When the occurrence of the disease is not regular and is infrequent, it is termed as sporadic.

(ii) Endemic

When the presence of the disease is constant in a particular geographical area, it is known as endemic. Endemic turns into a hyperendemic situation when a high level of disease occurrence is observed.

(iii) Epidemic

When there is a sudden rise in the number of patients with the same disease and within a particular area, it is termed as an epidemic.

(iv) Pandemic

When epidemics affect larger geographical areas (including multiple countries and continents), it is known as a pandemic. [7]

A disease takes the form of an epidemic when the following two conditions are met. First is when several people are affected by an illness/disease that has a similar nature to the disease and the same root cause, and the second is when the number of people infected increases rapidly over a period of time. When the epidemic crosses local boundaries and at the same time affects a large geological area, it becomes a pandemic. [8]

Another term that is primarily used when studying infectious diseases is “outbreak.” An

outbreak occurs when a sudden increase in the number of patients is observed. Outbreaks can last a few days, weeks, or months. A pandemic is also sometimes referred to as an outbreak. [9]

2.2.2 History of Pandemics

Some pandemics stand out in history because of the catastrophe they have caused.

FIGURE 2. Close to 25 million people are believed to have been wiped out by the early 1950s, which was approximately one-third of the European continent. [10]

- *Notable pandemics before 1800*

  The first and one of the worst pandemics witnessed by the world was known as the bubonic plague, also known as the Black Death Pandemic, in 1347. Millions of people lost their lives in the wave of this pandemic. [11] *At the beginning of 1500, the world experienced the smallpox pandemic. The death rate in some communities was 50%. This pandemic devastated many indigenous societies. [12] In 1881 the Fifth cholera pandemic occurred. More than 1.5 million deaths were reported. [13]*

- *Notable pandemics in 1900*

  In the early 1900, the Spanish flu influenza pandemic occurred. Fifty to hundred million deaths were reported [14]. In 1950 the Asian flu influenza happened. A total of 1.5 million deaths were reported. [15] In 1968 the Hong Kong flu influenza pandemic occurred. A total of 1 million deaths were reported. [16] Finally in 1981, the HIV/AIDS pandemic occurred which claimed 36.7 million deaths. [17] These pandemics caused a major economy loss. [18][19]

- *Pandemics after 2000.*

In the 2000s there was a whole new wave of pandemics. Severe acute respirator syndrome (SARS) and the Middle East respiratory syndrome (MERS) were viral diseases. [20] SARS occurred in 2003 which claimed 744 lives [21]. MERS occurred

in 2012 which claimed 659 deaths [22]. In 2009 there was the Swine flu influenza pandemic. It was also known as H1N1. This virus claimed 575,500 lives all over the world [23]. In 2013 the West Africa Ebola virus pandemic caused 11,323 deaths [24].

Historical records indicate that these pandemics pose a serious threat to humanity. The latest pandemic is Coronavirus Disease 2019 (COVID-19). COVID-19 was declared a pandemic by the World Health Organization (WHO) in January 2020. In a very short space of time, this pandemic has struck a large geographic area. [25]

2.2.3 The novel coronavirus

The word "novel" means "unknown" or "dealing with something new". From the beginning of the outbreak, extensive efforts are being demonstrated by scientists and experts all over the world. These extensive efforts include detection of the source of COVID-19, transmission pattern of the virus/pathogen, risk factors, disease progression, healthcare management, etc.

COVID-19 has a zoonotic origin. It means this virus was transmitted from animals to humans. The WHO has confirmed that COVID-19 can spread through air droplets. If the droplet produced by the infected patient is inhaled by a healthy person, then the healthy person can contract the infection. Symptoms of COVID-19 range from a person showing no signs (asymptomatic) to a person having a severe case of pneumonia. It has been observed that there are some recorded deaths of young people who after getting affected quickly succumbed to death. Bats are believed to be the source of COVID-19. However, the main animal source for COVID-19 has not been identified to date. The identification of the agent has not yet taken place either. In theory, the agent may be responsible for transmitting the virus from animals to humans.

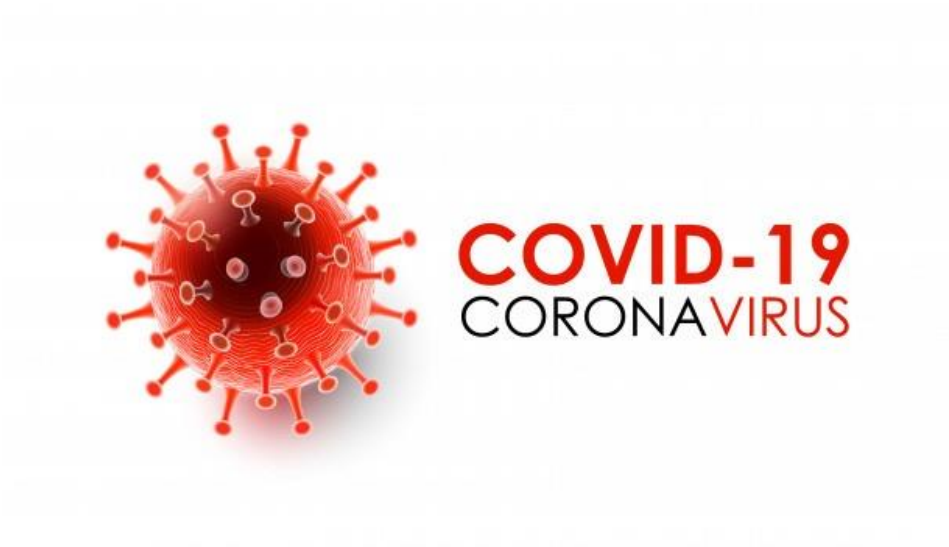


FIGURE 3. Coronavirus COVID-19. [26]

As the number of patients began to rise, it was clear that there was significant human-to-human transmission, but the pattern of transmission and the spread of pathogenesis in humans are still a mystery. It is also a big question whether the pathogenesis of the virus is increased or decreased over time. It is also a big question whether the pathogenesis of the virus will increase or decrease over time. Eventually, if the transmission rate drops, the spread of the disease will stop and the outbreak will end. If the transmission rate continues to rise, the community outbreak will go beyond the point of management. Since some patients have mild to no symptoms, it becomes very difficult to identify them. Asymptomatic infection can be very fatal in children.

The COVID-19 outbreak was an unprecedented situation that no one saw coming. The situation around COVID-19 is quickly becoming chaotic as the number of patients increases worldwide. Not only the number of those infected, but also the number of those who have died is increasing exponentially. Countries are applying the best possible control measures to contain the spread. Some of the countries have had little success in controlling the COVID-19 situation. However, there are numerous secrets surrounding the disease, starting with the origin itself. In this situation, professionals from different disciplines have to work together to find a solution. [27]

3 Exploratory data analysis (EDA)

Predicting mortality and the rate of spread plays a very important role in measures to control pandemic diseases such as COVID-19. Based on this prediction, preventive action can be taken by the public, government and health systems. These predictions are also helpful for pharmaceutical companies to help formulate and manufacture drugs faster. There are several techniques and models available to predict the spread / death rate. [28] This prediction is made based on the data available for the prediction. For pandemic diseases, researchers pull data from different data sources and use different models to analyze the data. The data can be referred from the following data sources:

- *John Hopkins University [29][30]*
- *Our World in Data [31]*
- *Google [32]*

The authenticity of the data source is controversial as these data sources are not approved by any standardization authority / agency; however, most of these data sources are nationalized repositories and WHO repositories. Some data can be in a structured format while others can be in an unstructured or semi-structured format. This heterogeneity of data is an important issue in data analysis. Analysis of various data sources and forecasting techniques can be useful for model selection [33].

This project focuses on analyzing data from July 1, 2020 to May 31, 2021 of three groups:

(i) Global
(ii) Three countries with highest confirmed cases: United States, Brazil, India
(iii) Three Asian countries: Japan, South Korea, Vietnam

First, we examine the number of confirmed, deaths, and recovered people in the countries studied. Then do calculation that focuses on three aspects of the pandemic:

1. Case-fatality rate
2. Recovery rate
3. Mortality rate

In addition to comparing such aspects, our analysis also takes into account factors such as Population, Age group (of confirmed cases) and Stringency Index.

3.1 Data retrieve

3.1.1 Libraries and tools

All of the code was developed in Python. [34] We mainly used Pandas library for retrieving, storing and manipulating data. [35] For visualization, we used Matplotlib and Seaborn library. [36][37]

Various computer softwares were also used to make working with data easier (Microsoft Office Excel 2016, IBM SPSS). [38][39]

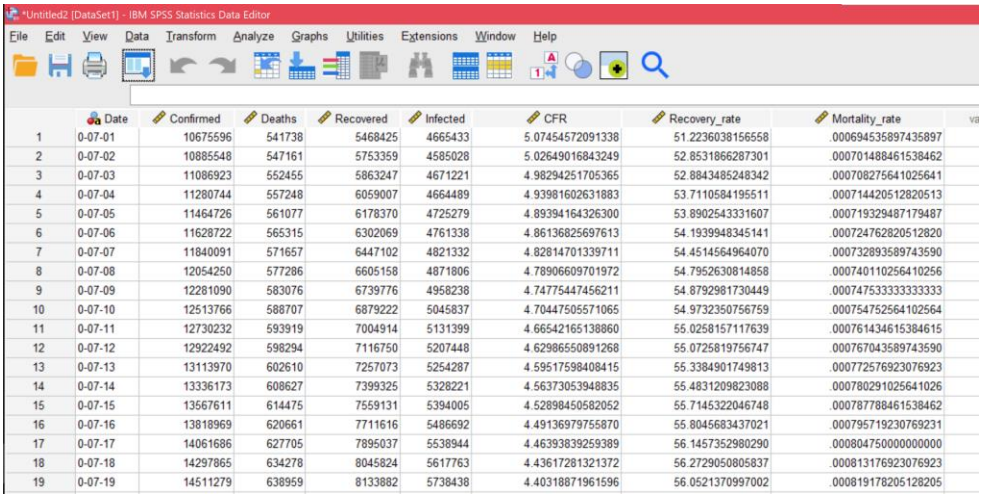
3.1.2 Data extraction results

3.1.2.1 Main datasets

There are a total of seven separated data tables for seven regions examined, each with eight attributes (column): *Date*, *Confirmed*, *Deaths*, *Recovered*, *Infected*, *Case-fatality rate (CFR)*, *Recovery rate* and *Mortality rate*, respectively. In which:

- ***Infected*** = Confirmed – Deaths – Recovered: Active cases
- ***Case-fatality rate*** = (Deaths / Confirmed) x 100: Rate of deaths over confirmed cases
- ***Recovery rate*** = (Recovered / Confirmed) x 100: Rate of recovered over confirmed cases
- ***Mortality rate*** = (Deaths / Population) x 100: Rate of deaths over a total population.

The time interval of this study ranges from July 1, 2020 to May 31, 2021. Each example (row) represents one case (day). There are a total 335 examples excluding the title row itself. Data of the United States contains missing records of Recovered from December 12, 2020 and are replaced with 0s.



	Date	Confirmed	Deaths	Recovered	Infected	CFR	Recovery_rate	Mortality_rate
1	0-07-01	10675596	541738	5468425	4665433	5.07454572091338	51.22360381565558	.00069453589743589
2	0-07-02	10885548	547161	5753359	4585028	5.02649016843249	52.8531866287301	.00070148846153846
3	0-07-03	11086923	552455	5863247	4671221	4.98294251705365	52.8843485248342	.00070827564102564
4	0-07-04	11280744	557248	6059007	4664489	4.93981602631883	53.71105841955511	.00071442051282051
5	0-07-05	11464726	561077	6178370	4725279	4.89394164326300	53.8902543331607	.00071932948717948
6	0-07-06	11628722	565315	6302069	4761338	4.86136825697613	54.1939948345141	.00072476282051282
7	0-07-07	11840091	571657	6447102	4821332	4.82814701339711	54.45145649640700	.00073289358974359
8	0-07-08	12054250	577286	6605158	4871806	4.78906609701972	54.7952630814858	.00074011025641025
9	0-07-09	12281090	583076	6739776	4958238	4.74775447456211	54.8792981730449	.00074753333333333
10	0-07-10	12513766	588707	6879222	5045837	4.70447505571065	54.9732350756759	.00075475256410256
11	0-07-11	12730232	593919	7004914	5131399	4.66542165138860	55.0258157117639	.00076143461538461
12	0-07-12	12922492	598294	7116750	5207448	4.62986550891268	55.0725819756747	.00076704358974359
13	0-07-13	13113970	602610	7257073	5254287	4.59517598408415	55.3384901749813	.00077257692307692
14	0-07-14	13336173	608627	7399325	5328221	4.56373053948835	55.4831209823088	.00078029102564102
15	0-07-15	13567611	614475	7559131	5394005	4.52898450582052	55.7145322046748	.00078778461538462
16	0-07-16	13818969	620661	7711616	5486692	4.49136979755870	55.8045683437021	.00079571923076923
17	0-07-17	14061686	627705	7895037	5538944	4.46393839259389	56.1457352980290	.00080475000000000
18	0-07-18	14297865	634278	8045824	5617763	4.43617281321372	56.2729050805837	.00081317692307692
19	0-07-19	14511279	638959	8133882	5738438	4.40318871961596	56.0521370997002	.00081917820512820

FIGURE 4. Full data table for Global, with 8 attributes.

All attributes are numeric variables with the exception of Date (variable of type String). Date variables are modified in the format Y-mm-dd, for Y is the last number of the year. This is because date and time data format can be misinterpreted by data analysis software (for example: 01-02-03 can be interpreted as: February 1, 2003; January 2, 2003; February 3, 2001), which leads to errors in the operation and storage of the process.

3.1.2.2 Custom datasets for model training

To simplify model training process as much as possible, we also created some custom datasets that contain only the criteria necessary to run the models. For example, to work with ARIMA model, we made separated datasets for each country with only two columns: Date and New Cases.

3.1.2.3 More custom datasets

We also considered the following datasets for just a little further observation:

- Population of studied countries
- Distribution of cases by Age (only in South Korea and Japan)
- Stringency Index (state strict level) for each region.

3.2 Statistical analysis

3.2.1 Basic visuals

3.2.1.1 General situation on a global scale

Figure 5 shows the overall situation of the COVID-19 pandemic from July 1, 2020 to May 31, 2021 on a global scale.

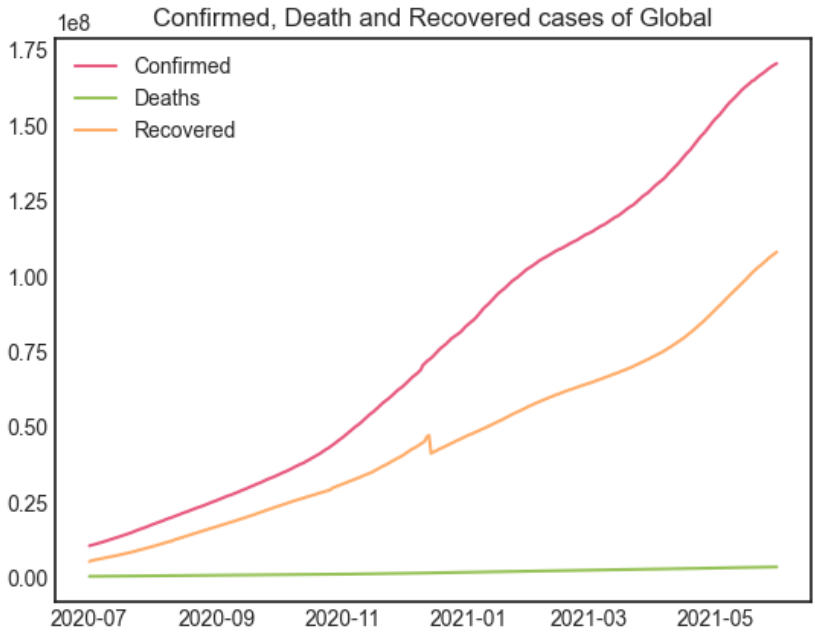


FIGURE 5. Confirmed, Death and Recovered cases of Global (until May 31, 2021).

As of May 31, 2021, there are over 170 million of cumulative confirmed cases reported, with more than 3 million people died. Recovery cases are around 108 million.

From previous observations on disease behavior, we assume that the actual Confirmed number is higher because there may be asymptomatic patients.

Recovered's line chart has a glitch towards the end of 2020. As noted above, since December 14, 2020, data is missing for the United States, along with some other countries that were not included in this study. It is probable that the US, which at the time has the largest number of confirmed cases in the world, was experiencing an unexpected outbreak making it difficult to keep track of cases.

3.2.1.2 Overall statistics of each country

The situation of COVID-19 from July 1, 2020 to May 31, 2021 of six countries is simulated in Figure 6.

The logarithm scale show percent change rather than the value of data. Since Deaths records are much lower than Confirmed and Recovered, we can get more information about the changes by using the logarithmic scale.

Except for United States, all figures show an upward trend that is almost linear. As mentioned, the US lacks data on the recovered cases as of December 14, 2020.

As the Confirmed case increases, the number of confirmed cases increases, so does the Deaths rate with a similar ratio. In other words, if the Confirmed case saw a sharp increase, the Death case would be expected to also experience a

drastic increase.

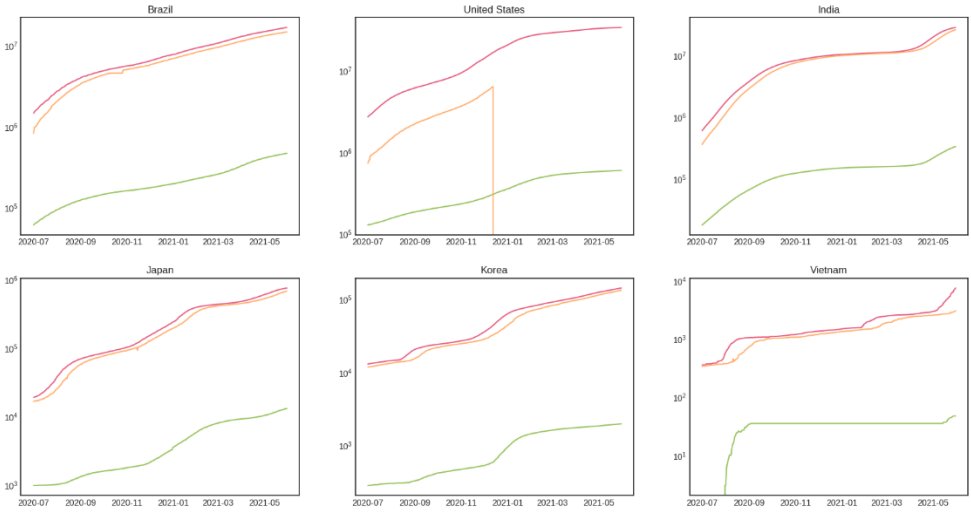


FIGURE 6. Confirmed, Death and Recovered cases: Brazil, United States, India, Japan, South Korea and Vietnam, respectively (until May 31, 2021) - in logarithm scale.

The Vietnam graph shows that the Confirmed cases were suddenly exploded from August 2020 to September 2020, causing Deaths cases to go up rapidly. Additionally, because there were not many previous Deaths cases in Vietnam, the Deaths line showed a strong exponential increase.

3.2.1.3 Daily confirmed cases of 6 countries

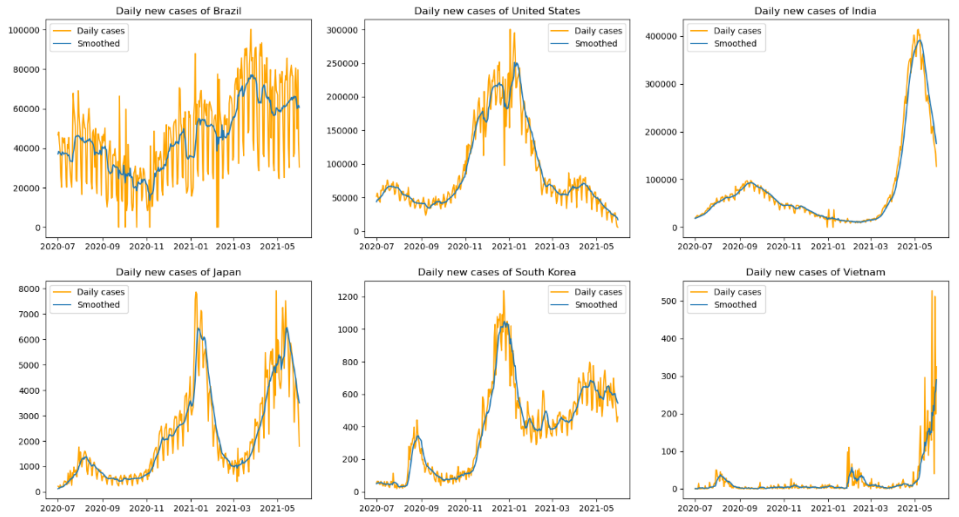


FIGURE 7. Daily confirmed cases: Brazil, United States, India, Japan, South Korea and Vietnam, respectively. Raw data is represented by orange, while smoothed data is graphed in blue.

Figure shows daily new cases reports of COVID-19 patients in adopted countries from July 1, 2020 to May 31, 2021. Here, the blue line shows a “smooth” behaviour as a result of data smoothing process.

Cumulative new cases will have always have the trend of exponential growth, while daily new cases vary each day.

Overall, it is expected that fewer cases occur every day in all countries as more and more people are being vaccinated. However, there was a sharp increase in some countries in April and May, particularly India, Japan and Vietnam. All of the countries mentioned are recovering as daily new cases began to decline from early May, with the exception of Vietnam, which is still increasing rapidly. In addition, Brazil is the country with the most diverse trend as it changes regularly.

In additional, the table below shows the date countries started mass vaccination:

Global	02/12/2020
US	14/12/2020
Brazil	17/01/2021
India	16/01/2021
Korea	26/02/2021
Japan	17/02/2021
Vietnam	08/03/2021

TABLE 1. Starting date of vaccination by country.

The US is one of the countries that is doing well with the vaccine. As of early 2021, new cases in the United States declined faster than ever, suggesting that vaccination progress in the country is going smoothly. However, the total numbers do not change significantly, even if vaccinations are already in place. The benefits are likely to come in the long run.

3.2.1.4 Daily fatal cases of 6 countries

Figure shows daily death reports of COVID-19 patients in adopted countries from July 1, 2020 to May 31, 2021. The United States faced an outbreak which was later over taken by Brazil along with India. Japan and Korea have also shown significant death rate over the period of time. Vietnam were able to control deaths to some extent using government forced lockdowns or following social distancing, etc. The virus has taken lives of many people worldwide.

The historical data depicts that the COVID-19 badly affected the countries which do not impose lockdowns or do not followed social distancing. Some variations in virus spread rate, recovery rate, and death rate can be seen in different countries based on population density, available health system in a country, testing capability, and action taken to contain the outbreak.

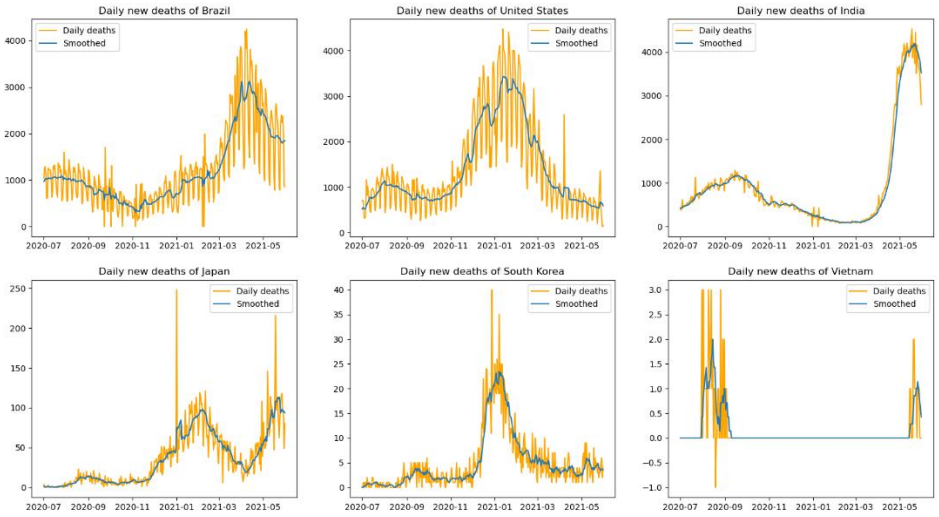


FIGURE 8. Daily death reports: Brazil, United States, India, Japan, South Korea and Vietnam, respectively. Raw data is represented by orange, while smoothed data is graphed in blue..

3.2.1.5 Case-fatality rate (CFR)

Figure 7 shows Case-fatality Rate (CFR) in terms of all regions.

The case-fatality rate is defined by

$$CFR = \frac{Deaths}{Confirmed} \times 100\ (\%)$$

which means the ratio of dead people to total infections.

With the exception of Vietnam, the graphs show an exponential decay trend. In addition, Japan's CFR decreases the most over time, suggesting that Japan is well under control of the disease.

There are some slight increases followed by decreases in South Korea and Japan line graphs. This might be due to some sudden outbreaks in these countries.

Before September of 2020, the Vietnam line chart was the only one that increased drastically while the others decreased. This is also when the Confirmed cases escalated, as mentioned. It is possible that the country was controlling the pandemic so well that its people lowered their guards, which allowed the disease to spread. However, from November 2020, it started to drop. And from February 2021, the line decreased rapidly, so Vietnam must have had the pandemic under control.

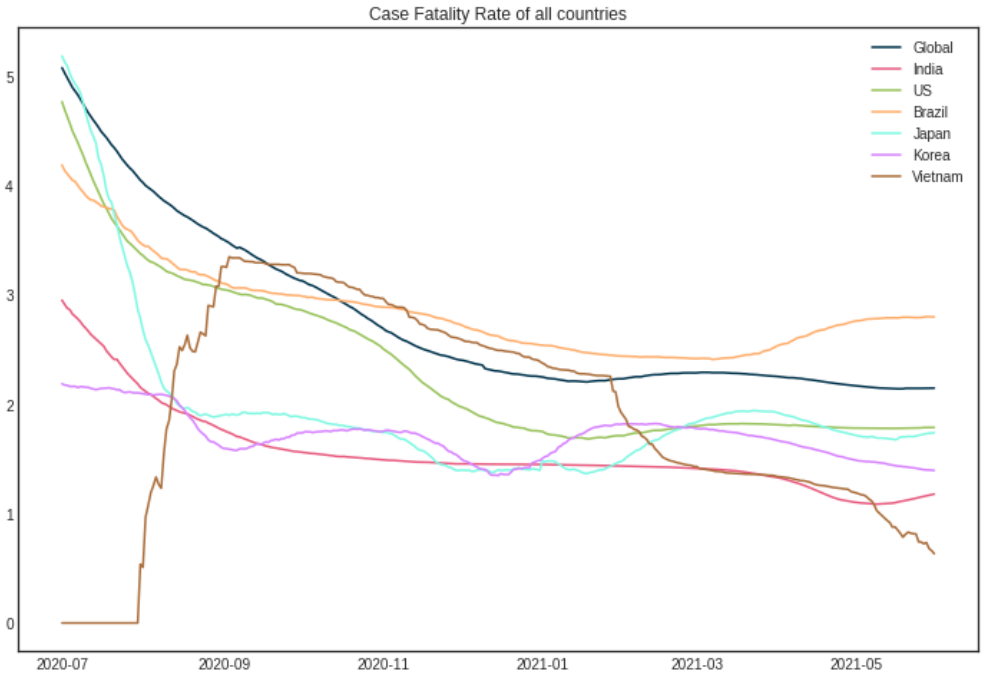


FIGURE 9. Case-fatality rate by region.

3.2.1.6 Recovery rate

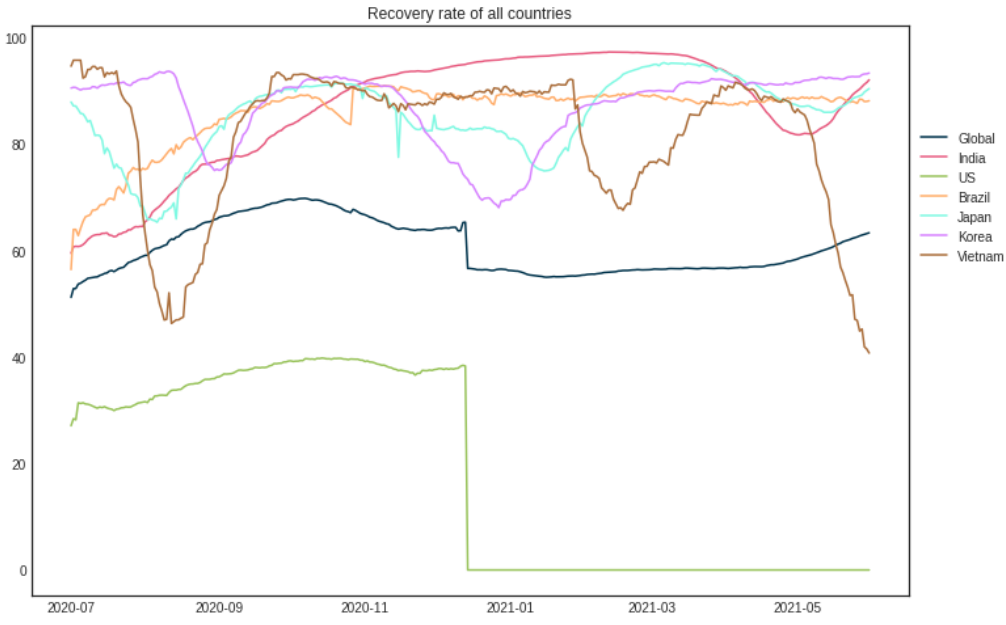


FIGURE 10. Recovery rate by region.

The recovery rate is defined by

$$Recovery\ rate = \frac{Recovered}{Confirmed} \times 100\ (\%)$$

which means the ratio of recovered people to total infections.

The general trend on this graph appears to be moving up and down around a stationary line. Line graph of the United States is missing from December 14, 2020 onwards. US is the country with the highest number of Confirmed as well as Deaths globally. Hence, the US Recovery data should have a huge impact on global Recovered data trend.

3.2.1.7 Mortality rate

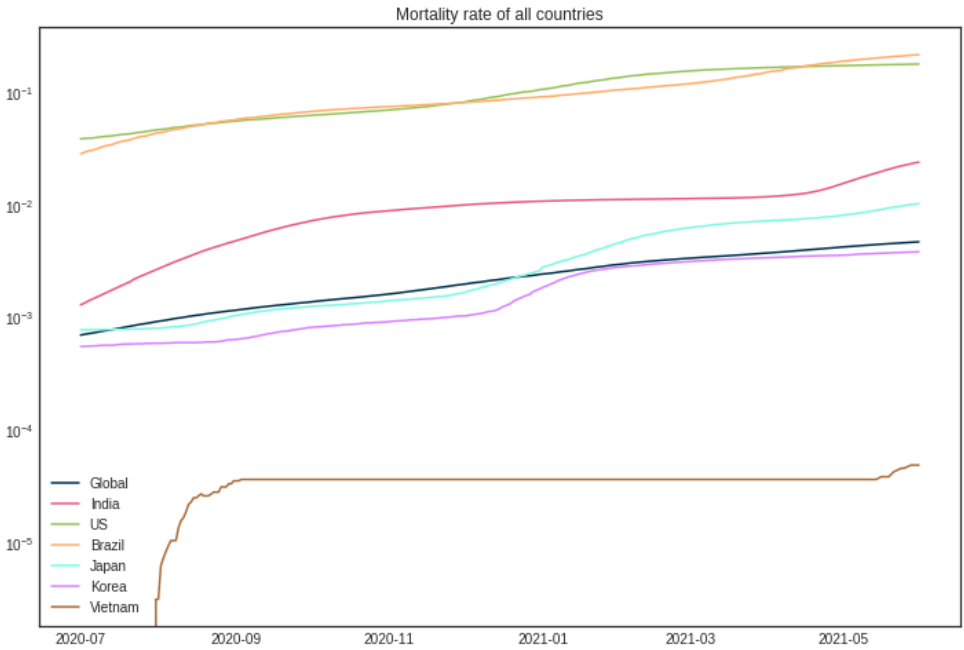


FIGURE 11. Mortality rate by region, on logarithm scale.

The mortality rate is defined by

$$Mortality\ rate = \frac{Deaths}{Population} \times 100\ (\%)$$

which means the ratio of dead people to total population of an area.

The graph shows that the rate is growing relatively smoothly, with the exception of Vietnam, which saw a sharp increase. Even if only a small proportion of the population (0.01% - 0.2%) is lower than the mortality rate of people who die from other reasons (traffic accidents, other diseases), we still have to take into account that

not all COVID-19 cases are recorded according to the observations in the previous sections. As a result, it is still a dangerous disease and if we don't make efforts to control it, the death rate could rise exponentially to the point of loss of control.

3.2.2 Histograms and distributions

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions. Although histograms are made up of bars and looks very much like bar charts, they are not the same. There are important differences:

- *Histogram* refers to a graph that displays data in the form of bars to indicate the frequency of numerical data; whereas *bar graph* is a graphical representation of data that uses bars to compare different categories of data.
- *Histogram* is used for nondiscrete variable distribution, while *bar graph* is used for discrete variables comparison.
- With *bar charts*, it is common to rearrange the blocks, from highest to lowest. This is not possible with *histograms*.

In conclusion, histogram is used to show the frequency of occurrences and bar graphs are useful for comparing different categories of data. [40]

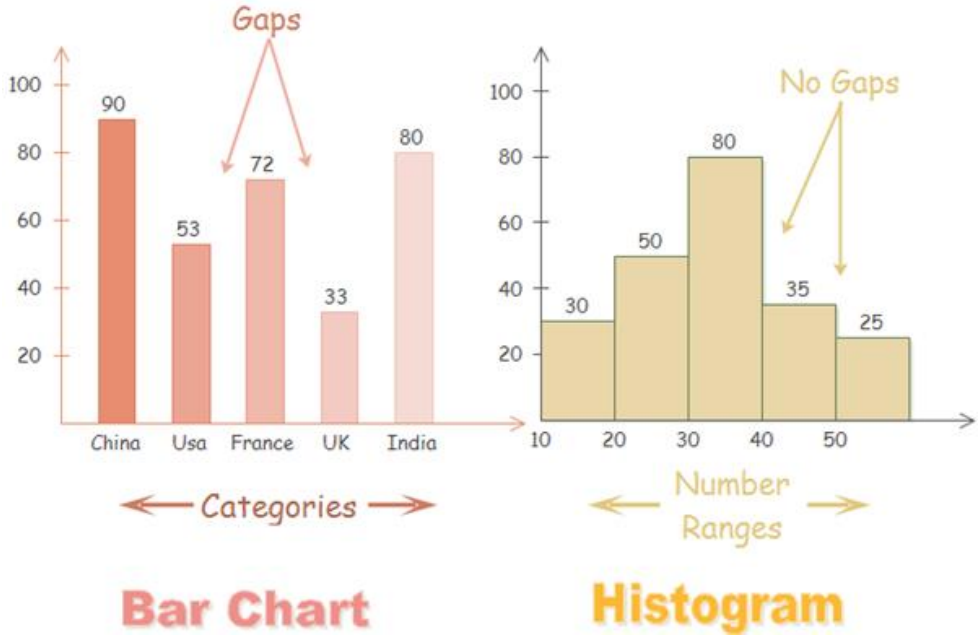


FIGURE 12. Differences between bar chart and histogram. [41]

Some of the most common types of histogram distribution are listed below:

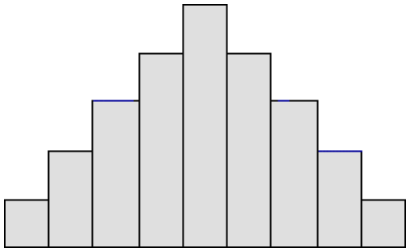


FIGURE 13. Normal distribution.

A common pattern is the bell-shaped curve known as the *normal distribution*. In a normal or "typical" distribution, points are as likely to occur on one side of the average as on the other. There are other distributions that look very similar to the normal distribution. Statistical calculations must be utilized to prove a normal distribution.

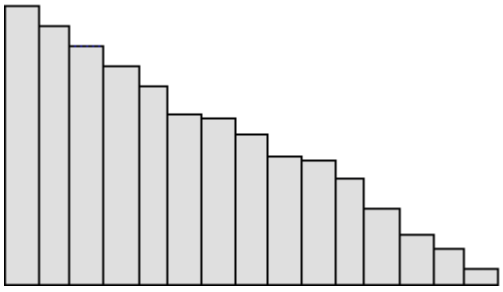


FIGURE 14. Right-skewed distribution.

The skewed distribution is asymmetrical because a natural boundary prevents results one side. The apex of the distribution is eccentric towards the border and a tail extends from it. For example, a distribution of analyses of a very pure product would be skewed, because the product cannot be more than 100 percent pure. Other examples of natural limits are holes that cannot be smaller than the diameter of the drill bit or call-handling times that cannot be smaller than zero. These distributions are called right- or left-skewed depending on the direction of the tail.

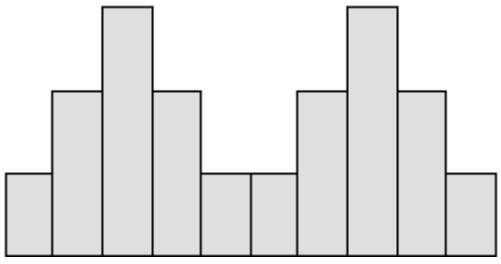


FIGURE 15. Bimodal (double-peaked) distribution.

A bimodal shape has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.

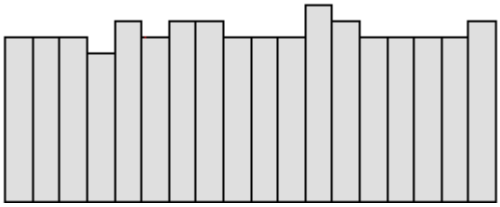


FIGURE 16. Uniform distribution.

A uniform distribution provides little information about the system. An example would be a state lottery where each class has roughly the same number of elements. A uniform distribution often means that the number of classes is too small.

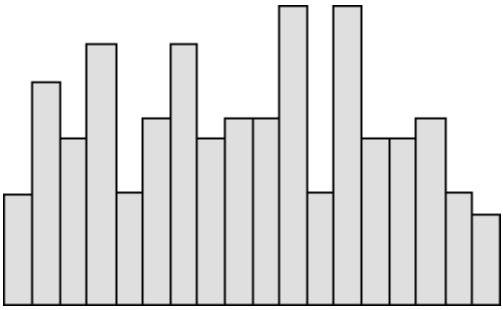


FIGURE 17. Random distribution.

A random distribution has no obvious pattern. Like the uniform distribution, it may describe a distribution that has several modes (peaks). A random distribution often means there are too many classes. [42][43]

Mathematical distribution such as Laplace distribution, Gaussian distribution, Poisson distribution,... are also often used in statistical analysis.

3.2.3 Data histograms

We observe the histograms created from recorded data of each region.

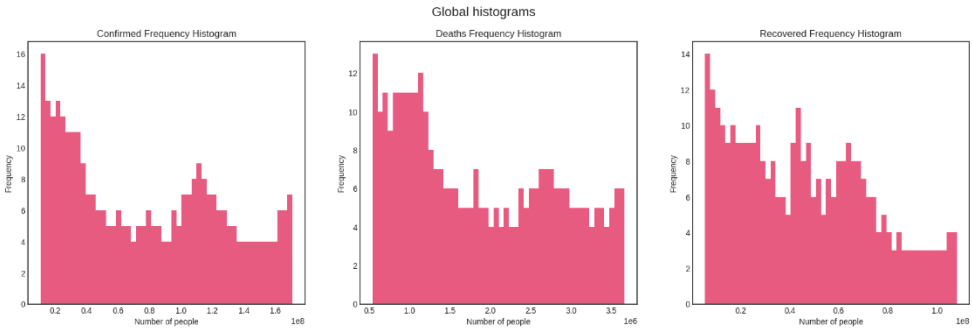


FIGURE 18. Global Confirmed, Death and Recovered frequency histograms. The overall shapes of three figures appear somewhat **skewed to the right** with several peaks (**multimodal**). This means the data tends to have lower values.

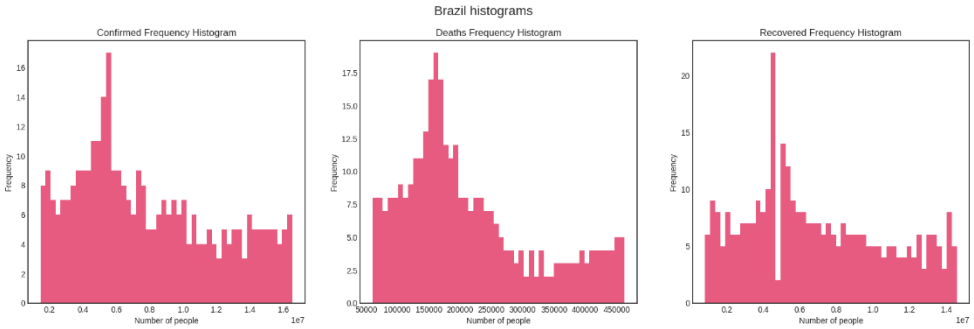


FIGURE 19. Confirmed, Death and Recovered frequency histograms of Brazil. All three graphs show the same trend of a **right-skewed distribution**.

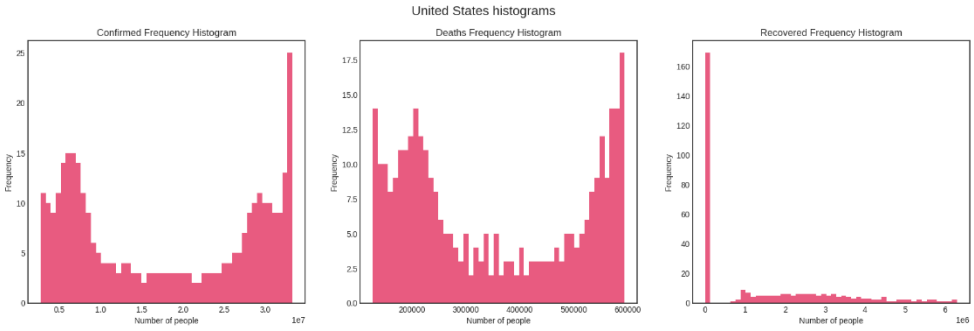


FIGURE 20. Confirmed, Death and Recovered frequency histograms of United States. The last graph seems **randomly distributed**, while the other two somewhat follow the ***dog food distribution*** – meaning the data tends to split into two clusters at either ends, creating a large gap in between remains.

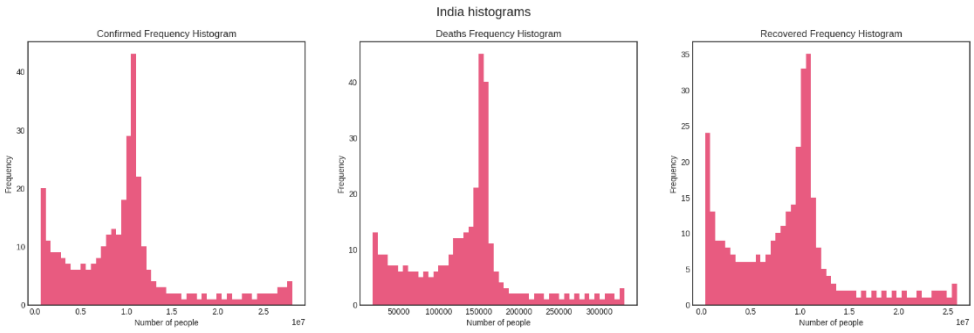


FIGURE 21. Confirmed, Death and Recovered frequency histograms of India. The three plots closely follow the **Laplace distribution** with a large peak at the left end (edge peak).

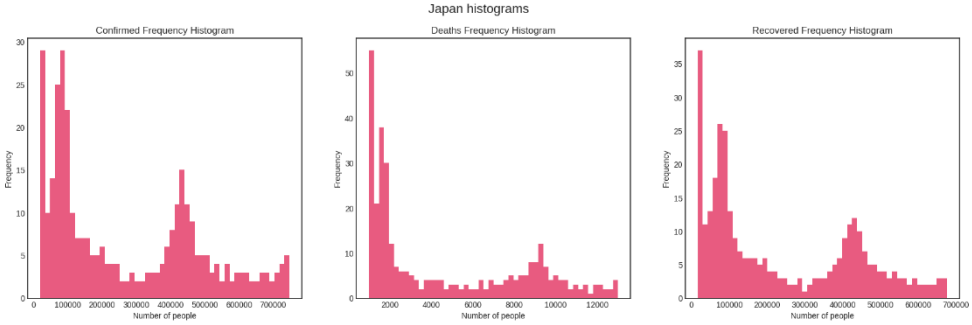


FIGURE 22. Confirmed, Death and Recovered frequency histograms of Japan. All graphs show **multimodal right-skewed distribution**.

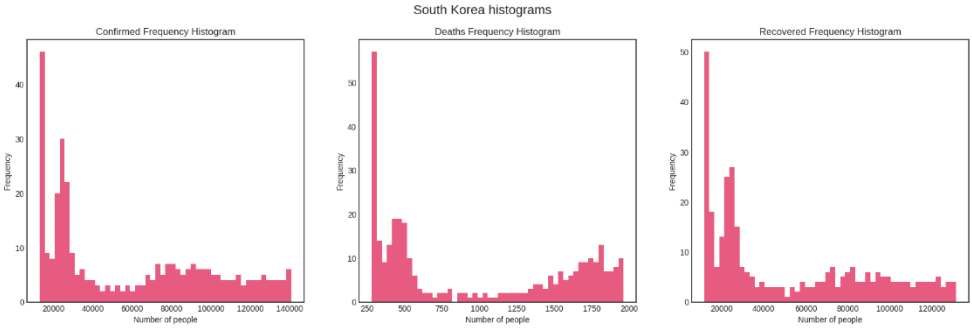


FIGURE 23. Confirmed, Death and Recovered frequency histograms of South Korea. The first and last graphs have the same shape of a **right-skewed trend** with a small elevation, while Deaths histogram looks like the **dog food distribution**.

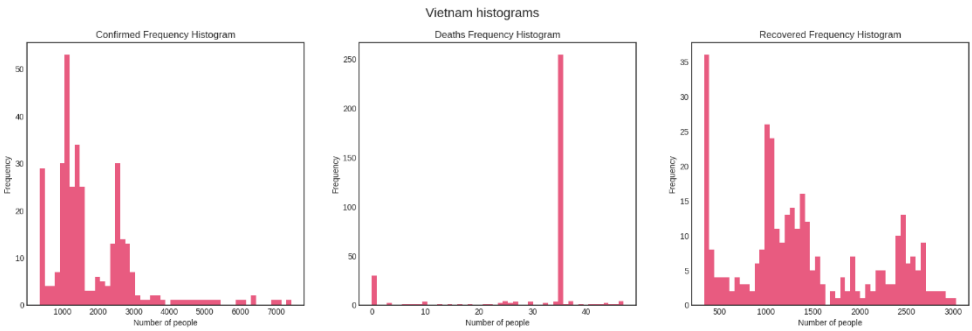


FIGURE 24. Confirmed, Death and Recovered frequency histograms of Vietnam. Data appear as **random distribution** as Vietnam has little information compared to others countries.

3.2.4 Custom visuals

In this section, we examine some other factors behind the pandemic.

3.2.4.1 Country population

India is the world's second most populous and has the highest population amongst all studied countries. In 2020, the estimated total population in India amounted to approximately 1.38 billion people, with average density of 464 people per km^2. [44]

India's current COVID-19 surge is an unprecedented public health crisis. Previous data analysis in Section shows that the country's confirmed daily cases are plateauing. But in reality, the true scale of the COVID-19 outbreak in India is impossible to accurately quantify.

The populations of Japan, Korea, and Vietnam are relatively small. These are countries that have fewer confirmed cases compared to the other countries.

Studies show that population density facilitates transmission of disease via close person-to-person contact and may support sustained disease transmission due to increased contact rates. [45] Based on these observations, we see that population density levels at the county level can increasingly explain the variation of cumulative cases across counties as the epidemic progressed.

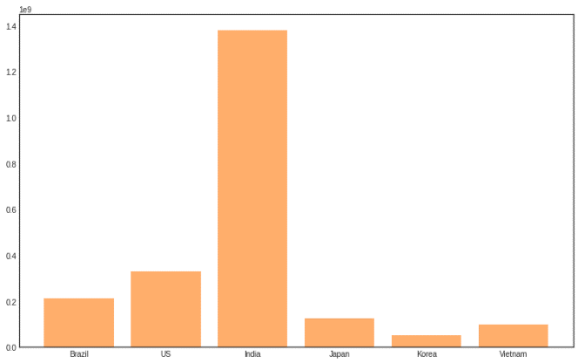


FIGURE 25. Population of studied countries.

3.2.4.2 Age distribution

Below are Japan and South Korea's confirmed cases histograms, distributed by age.

The graph of South Korea distributed normally. Most of the confirmed cases of Korea range from 20 to 59, while Japan has a surprisingly high number of 20s people diagnosed. The distribution of Japan is right-skewed.

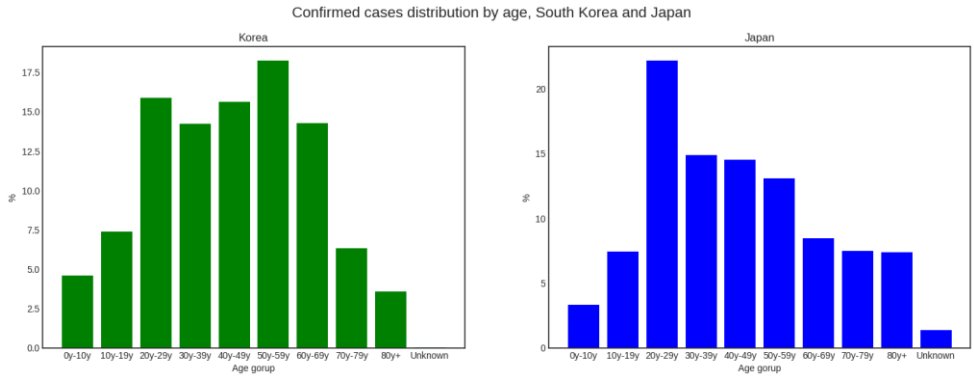


FIGURE 26. Confirmed cases distribution by age of South Korea and Japan.

It is observed that the reasons behind Japan's case are:

> *(i). Social: 20s tend to have more social activities than the others, for example: universities, dating, and hanging out.*
>
> *(ii). Work: Younger people tend to be delayed from school while universities continue to work.*
>
> *(ii). Romance Factor: The 20s are biologically the high point of man's romantic seasons, which means more close contact and activity.*
>
> *(iii). Psychological factors: COVID-19 is extremely dangerous for the elderly, but not so much for the younger ones. To most people, it just seems like a different type of flu. This could make them take containment measures (like quarantine and social distancing) less seriously. [46]*

While in the case of Korea, the most popular working age is 20-59. Participating in daily activities may put people at a higher risk of developing serious routes of infection.

3.2.4.2 Stringency Index

Stringency index is a composite score developed by researchers at Oxford University to compare countries' policy responses to the COVID-19 pandemic.

Each country's lockdown is different. The wide range of actions taken by different governments presents a challenge for analysts who wish to compare these guidelines over time or between countries. To enable such comparisons, a team at Oxford University's Blavatnik School of Government maintains a database of pandemic-response measures and derives an index of the measures' overall stringency. Public information on government response measures are collected and assigned stringency ratings which are then used to derive a composite score between 0 and 100.

According to the Oxford team, the index does not perfectly capture local measures in large or federal countries. A measure only in force in one or two regions contributes less to the stringency index than a nationwide policy, but rules in force in only one or two regions can also inflate a whole country's overall score. [47][48]

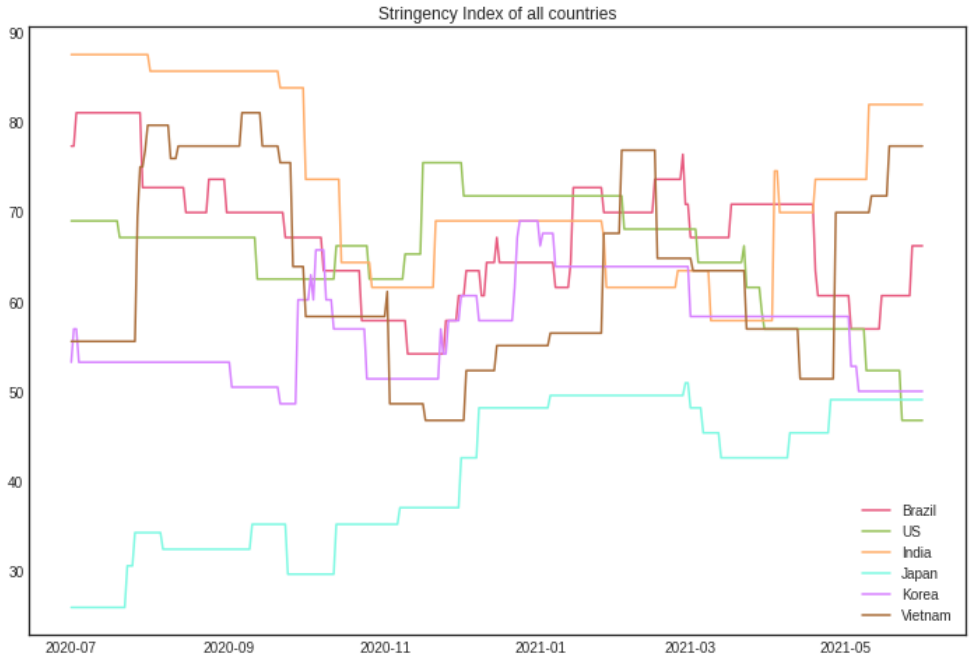


FIGURE 27. Stringency Index chart by country.

As can be seen from the graphic and the previous sections, countries with the higher case numbers tend to have a higher stringency index. The response of governments to the COVID-19 situation is important to keep the pandemic under control.

4 SIR model terminology 101

To understand the key elements affecting the transmission of COVID-19, we consider developing a predictive model. One of the most well-known models in epidemiology is SIR.

4.1 From basic exponential growth to the famous curves set

A person who carries the virus is said to be ***Infected***, and others who are not infected but still have a percentage chance of getting the virus spread by the *Infected* are said to be ***Susceptible***. Infectious people can turn susceptible people into even more infectious people. [49]

In a very simple scenario, for example, assume that each *Infected* can infect two *Susceptible*, which means that after a certain period of time, 1 *Infected* becomes 2, 2 becomes 4, 4 becomes 8, and so on; starting with only 1 of the population being *Infected* and without additional factors. The infection diagram is similar to the following figure [50]:

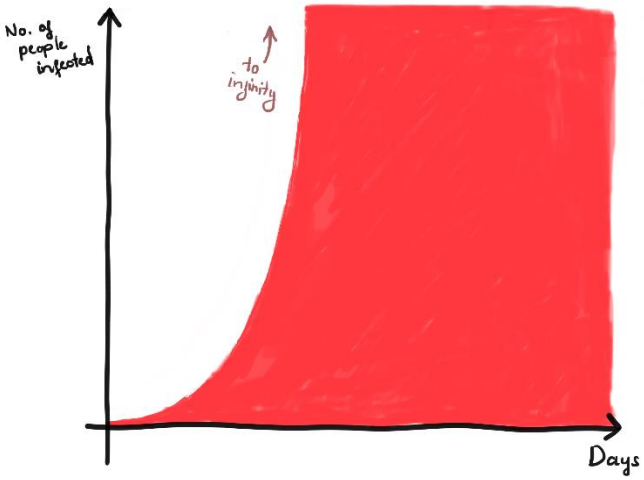


FIGURE 28: Growth of the number of *Infected* in the simplest scenario.

This trend follows the **basic exponential growth**: starts low and then unexpectedly blows up explodes to infinity.

Fortunately, that definitely doesn't happen in real life. The human population is a fixed number. Therefore, the trend cannot go on forever. The idea is:

- *When there is only one Infected and the others are Susceptible, the Infected can infect 100% of contacts*
- *When half of the population are Infected and another half are Susceptible, the Infected can infect 50% of contacts*
- *When all of the population are Infected, they cannot infect anyone else.*

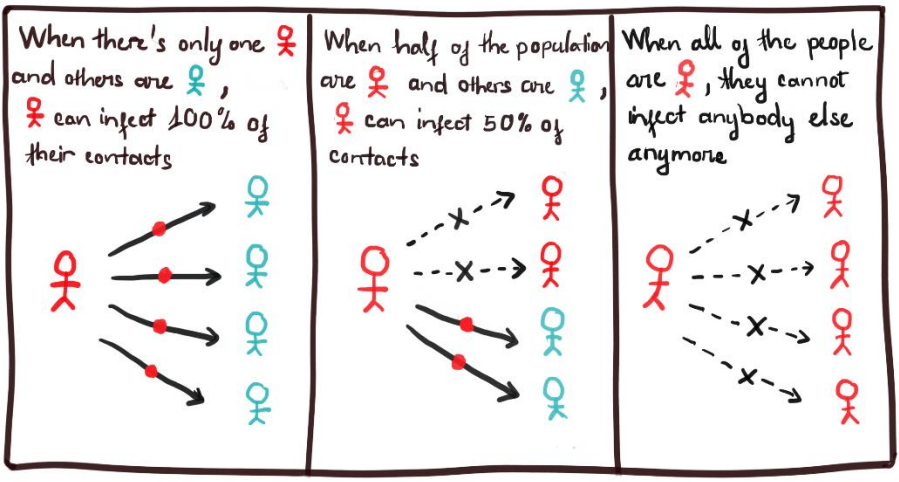


FIGURE 29: Possibility of virus transmission, simply explained.

In that sense, the more *Infected* there are, the faster *Susceptible* become *Infected;* but, the fewer *Susceptible* there are, the slower *Susceptible* become *Infected.* The epidemic growth graph has now been changed [51]:

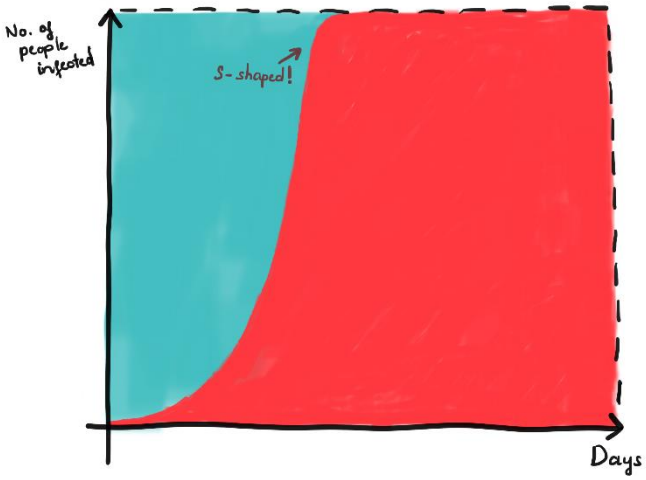


FIGURE 30: Growth of the number of *Infected* when the population capacity is fixed. The *Infected* is graphed in red while *Susceptible* is represented by the color teal.

This is called the **logistic growth curve**: starts low, explodes, and then slows down again.

However, the concept is not yet applicable. In reality, when the *Infected* stops being contagious, this means: either they are 1/ recovered (and healthy), 2/ immune to the virus but left with permanent disabilities, or in the worst case, 3/ dead. To put it simply, we pretend that all *Infected* become ***Recovered*** after some days and never be infected again. Assume that the population is fixed and contains only *Infected* and *Recovered*. Change in amount of each aspect is simulated below [52]:

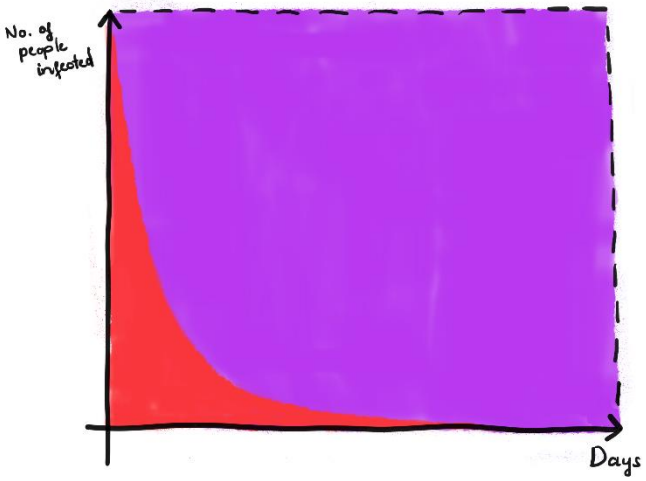


FIGURE 31: Graph of *Infected* assuming that all *Infected* will become *Recovered.* The *Infected* is graphed in red while *Recovered* is represented by the color purple.

This is the opposite of the above exponential growth - the **exponential decay curve**.

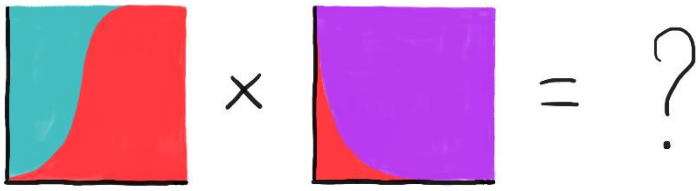


FIGURE 32: What would be the result if these scenarios happen at the same time?

Combine the S-shaped logistic growth of *Infected-Susceptible* with the decay curve of *Infected-Recovery*, we obtain the final result:

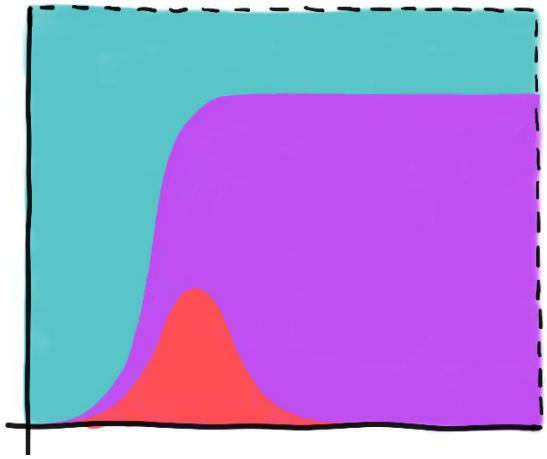


FIGURE 33: Graph of *Infected*, *Recovered* and *Susceptible* with respect to time (x-axis).

This set of curves is an important and common mathematical concept in data analytics - the SIR Model.

4.2 SIR Model

4.2.1 Theory

SIR model divides a fixed population of N people into 3 compartments:

- *Susceptible (S):* The individual has not contracted the disease, but can be infected due to transmission from infected people.
- *Infected (I):* The individual has contracted the disease.
- *Recovered (R):* The individual has survived and developed immunity to the disease or is deceased.

Consider that people develop immunity and there is no transition from recovered to the remaining 2 stages, the differential equations that govern the system are:

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

With

- beta (β): probability susceptible–infected contact results in a new exposure;
- gamma (γ): probability of one infected subject gets recovered;

This model requires as input the amount of the susceptible, infected, and cured or dead population, all referring to a reference time, called time 0. The parameters are necessary to establish the rates and the model dynamics.

The SIR model is one of the simplest compartmental models whose main purpose is to understand the key factors that impact the epidemic transmission, which, in this case, is β and γ. [53]

4.2.2 Results

The results for the SIR model are presented in the order established for the country analysis: Global, Brazil, USA, India, Japan, Korea and Vietnam.

With the aim of understanding the effects of the parameter selection for the SIR compartment model, a total of nine diagrams are presented for each country, taking into account a variation of the γ and β parameters. First, a fixed γ = 0.2, which corresponds to an infection duration of 5 days, is selected and the three different values for β are considered: β = [1, 2, 3]. This means that the proposed simulation will consider different possibilities for the capacity of the virus to spread through personal contacts and its inherent rate of transmission. The same process is then repeated for γ = 0.1, corresponding to an infection duration of 10 days and γ = 0.066, which corresponds to an infection duration of 15 days.

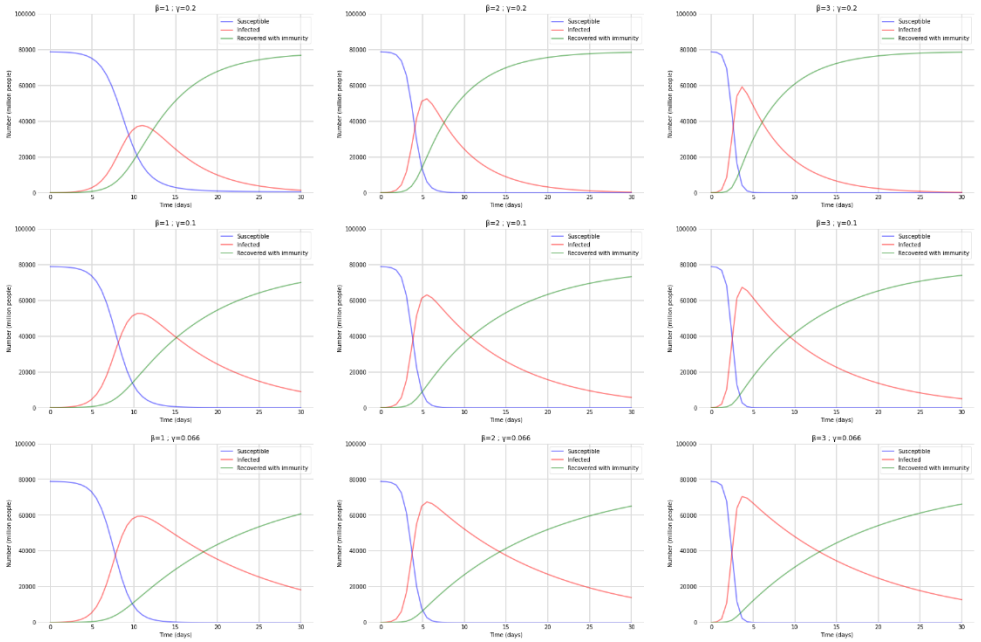


FIGURE 34. Results for Global: Prediction based on the SIR model considering the variation of the parameters γ and β. 1st, 2nd and 3rd figures has γ = 0.2 and β = [1, 2, 3], respectively. 4th, 5th and 6th figures has γ = 0.1 and β = [1, 2, 3], respectively. Figures 7th, 8th and 9th has γ = 0.066 and β = [1, 2, 3], respectively.

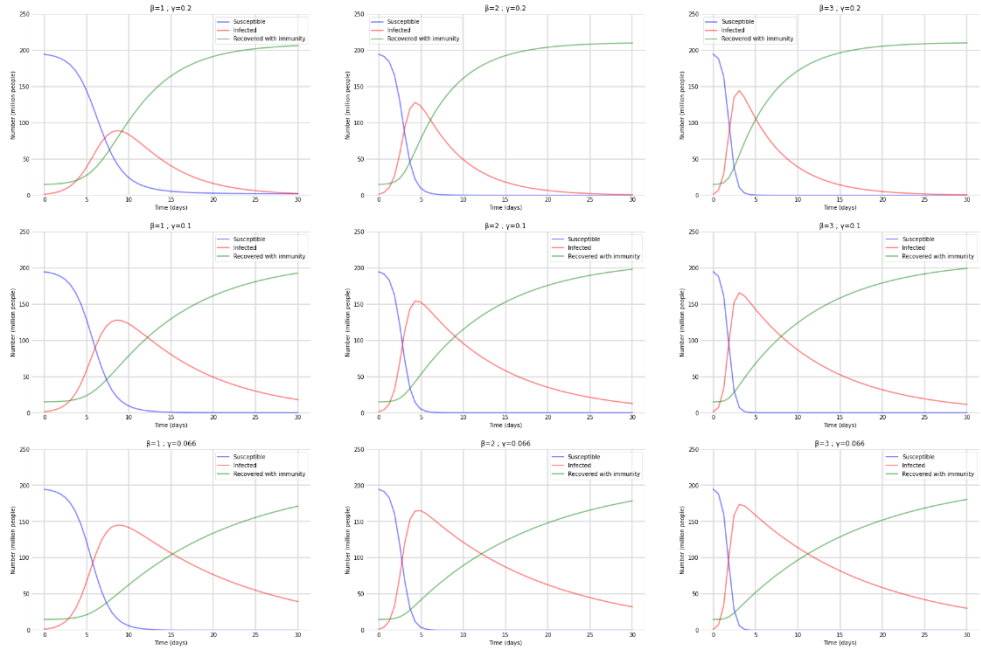


FIGURE 35. Results for Brazil: Prediction based on the SIR model considering the variation of the parameters γ and β. 1st, 2nd and 3rd figures has γ = 0.2 and β = [1, 2, 3], respectively. 4th, 5th and 6th figures has γ = 0.1 and β = [1, 2, 3], respectively. Figures 7th, 8th and 9th has γ = 0.066 and β = [1, 2, 3], respectively.

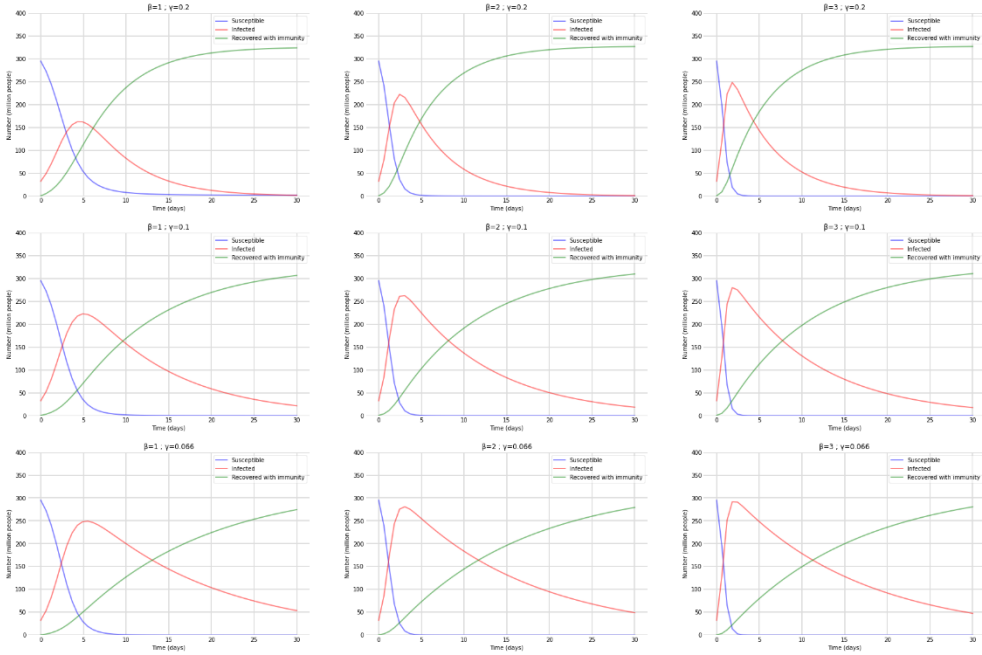


FIGURE 36. Results for United States: Prediction based on the SIR model considering the variation of the parameters γ and β. 1st, 2nd and 3rd figures has γ = 0.2 and β = [1, 2, 3], respectively. 4th, 5th and 6th figures has γ = 0.1 and β = [1, 2, 3], respectively. Figures 7th, 8th and 9th has γ = 0.066 and β = [1, 2, 3], respectively.

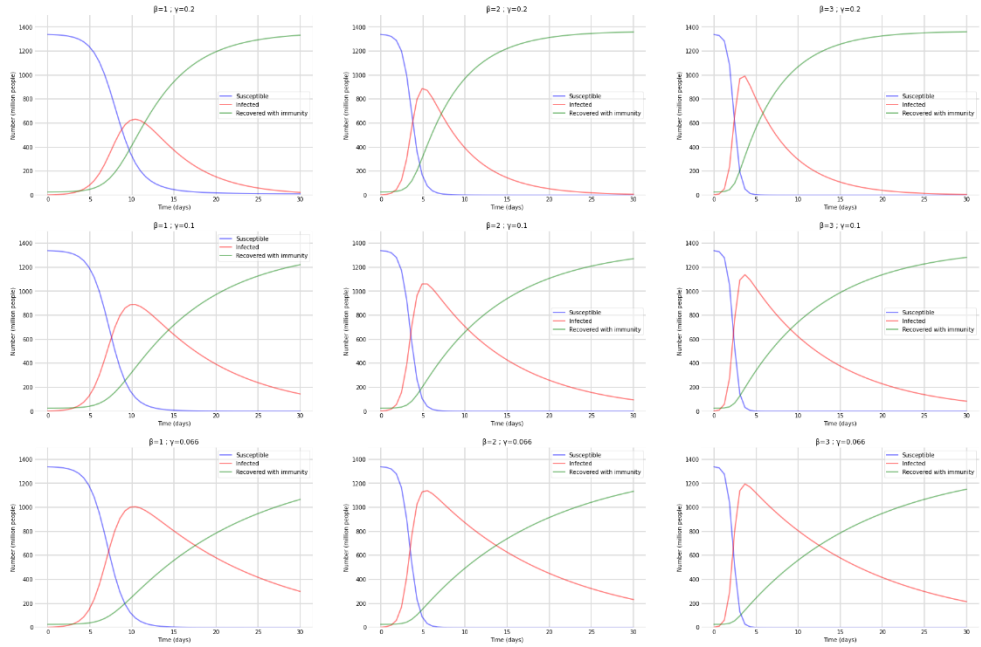


FIGURE 37. Results for India: Prediction based on the SIR model considering the variation of the parameters γ and β. 1st, 2nd and 3rd figures has γ = 0.2 and β = [1, 2, 3], respectively. 4th, 5th and 6th figures has γ = 0.1 and β = [1, 2, 3], respectively. Figures 7th, 8th and 9th has γ = 0.066 and β = [1, 2, 3], respectively.

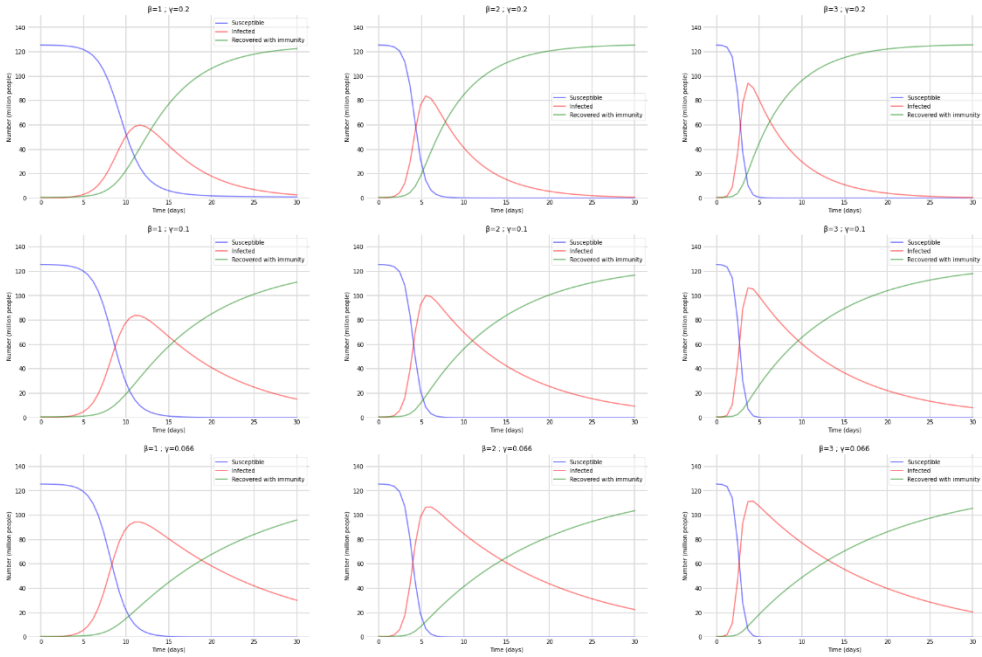


FIGURE 38. Results for Japan: Prediction based on the SIR model considering the variation of the parameters γ and β. 1st, 2nd and 3rd figures has γ = 0.2 and β = [1, 2, 3], respectively. 4th, 5th and 6th figures has γ = 0.1 and β = [1, 2, 3], respectively. Figures 7th, 8th and 9th has γ = 0.066 and β = [1, 2, 3], respectively.

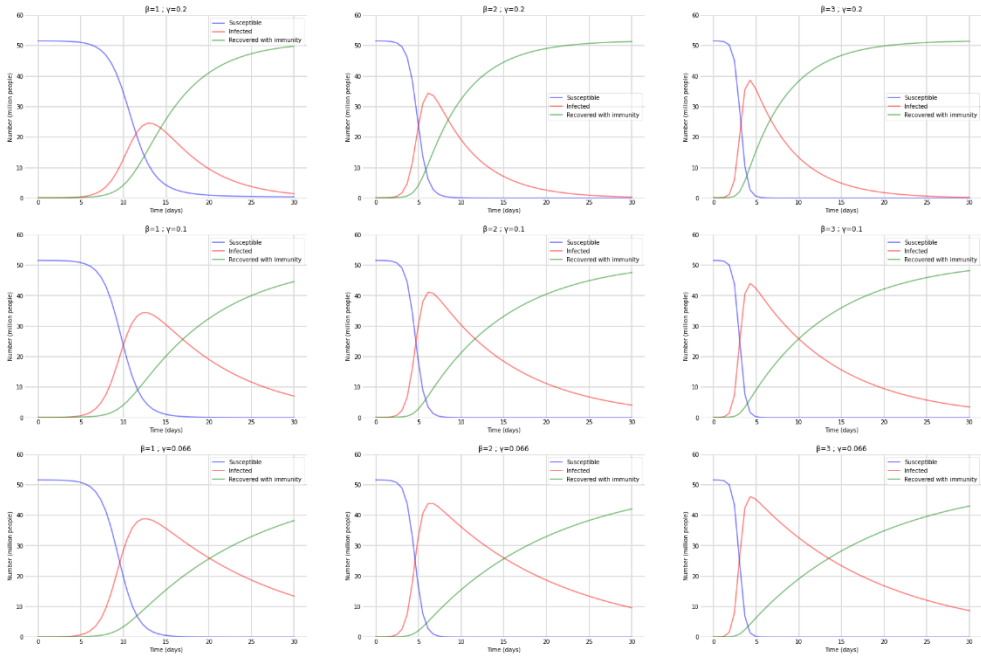


FIGURE 39. Results for South Korea: Prediction based on the SIR model considering the variation of the parameters γ and β. Figures a, b, and c has γ = 0.2 and β = [1, 2, 3], respectively. Figures d, e, and f has γ = 0.1 and β = [1, 2, 3], respectively. Figures g, h, and i γ = 0.066 and β = [1, 2, 3], respectively.

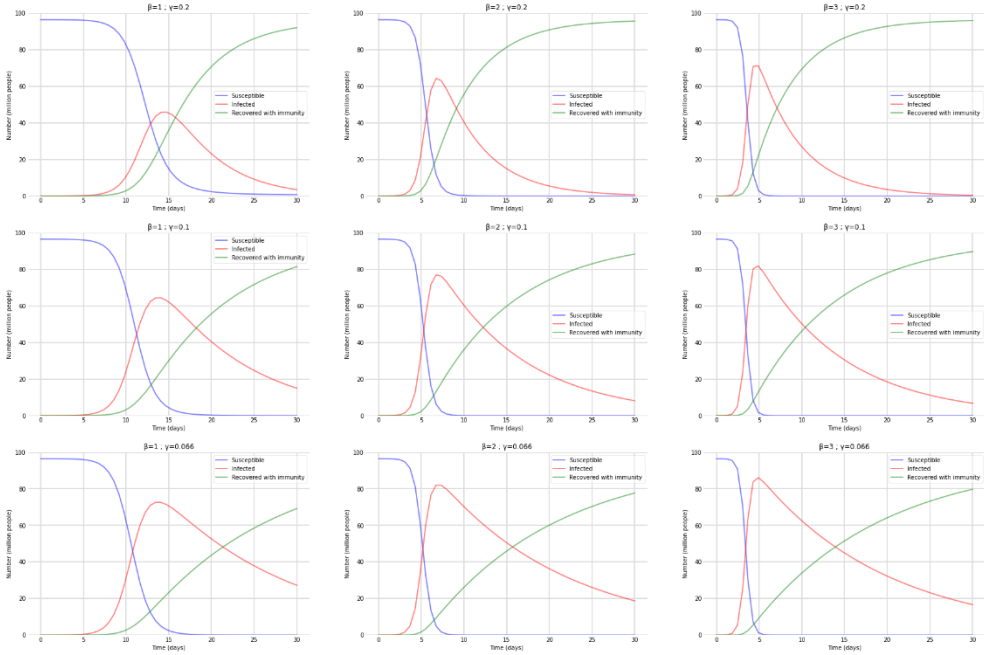


FIGURE 40. Results for Vietnam: Prediction based on the SIR model considering the variation of the parameters γ and β. Figures a, b, and c has γ = 0.2 and β = [1, 2, 3], respectively. Figures d, e, and f has γ = 0.1 and β = [1, 2, 3], respectively. Figures g, h, and i γ = 0.066 and β = [1, 2, 3], respectively.

4.2.3 Observations of the model

- *At the start of the period, the number of infected individuals increases if β > γ and vice versa.*
- *The number of susceptible individuals decreases over time.*
- *The number of infected individuals reaches a peak, then decreases over time. This occurs sooner as the parameter β increases.*
- *All compartments flatten out after a period of time, which occurs sooner as the parameter γ increases.*
- *The number of recovered or deceased individuals increases more or less significantly as the number of infected individuals increases or decreases respectively.*

The application of SIR is widely popular during the COVID-19 pandemic but many predictions weren't verified because the modeling could not represent the real models, which were dependent on several external factors. As expected, when using these models, the results presented for SIR was strongly dependent on the parameters selection. Each parameter is responsible for the rate of transitions between one compartment and the next one.

Compartmental models like SIR are valid approaches for comprehending and analyzing epidemiological data, especially if the model is adjusted to consider specific aspects of the

epidemic under analysis, as in the case of the COVID-19 pandemic. For example, at the beginning of the pandemic in some countries, several publications and articles were considering the model as ground truth and assuring that the pandemic would peak and exponentially decrease to near zero in just a few days, which has not been confirmed in real cases.

5 Machine learning in COVID-19 prediction

The aim of these models is to predict the number of infected/active cases or daily new cases of COVID-19.

5.1 Random Forest Regression (RFR) model

5.1.1 Theory

Random Forest Regression (RFR) is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. [54]

RFR splits the dataset into a number of trees (n_estimators [55]), then repeats the process in each tree for a number of times (max_depth [55]) or until all leaves are pure (all data points contain the same label). Then it takes in all of the output of the trees to form a final result.

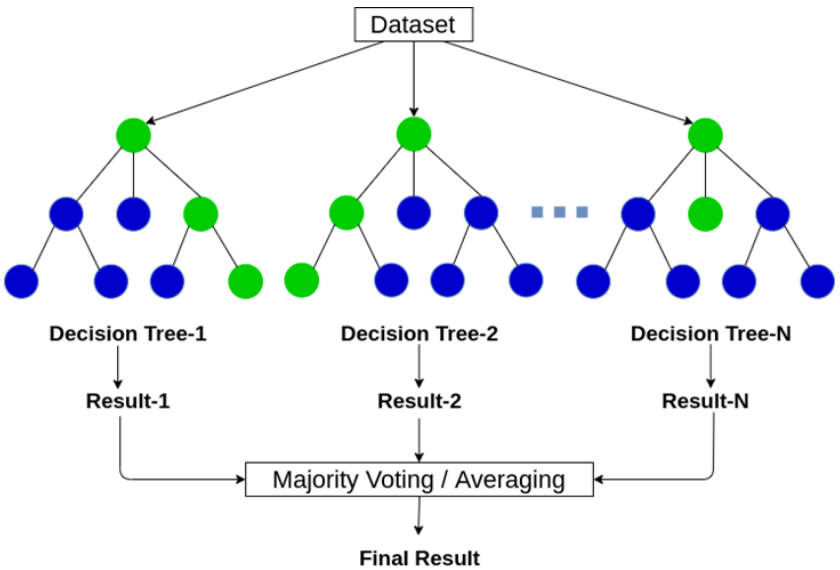


FIGURE 41. Random Forest model theory. [56]

5.1.2 Pre-proposed method

RFR requires an input value and an output value in order to function. First step is to convert our time-series dataset into a monitored learning dataset.

To predict the value for the last 14 days, the model will predict the next day's data from the values for the last known 6 days, then add the forecast to the data set and repeat the process for a total of 14 times.

5.1.3 Test results

The results were organized in the following sequence: Global, Brazil, United States, India, Japan, South Korea, Vietnam.

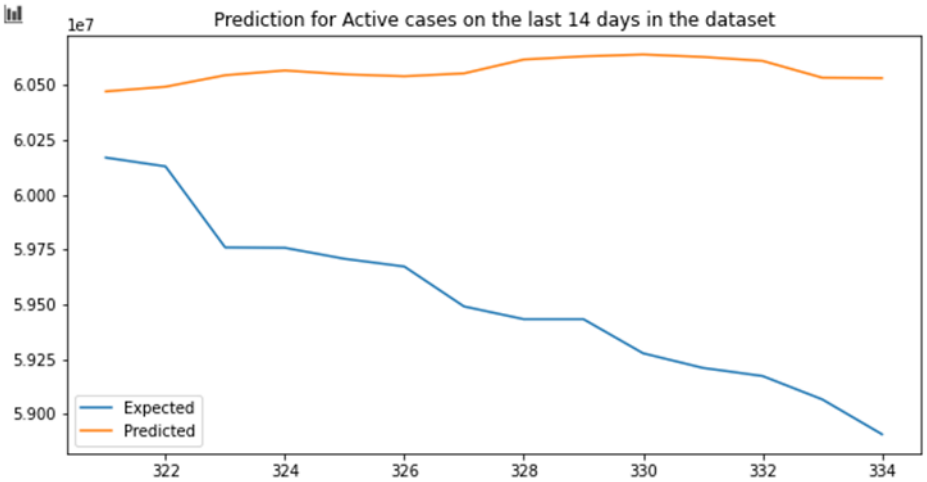


FIGURE 42. Test result of the RFR model for last 14 examples of Global data sample.

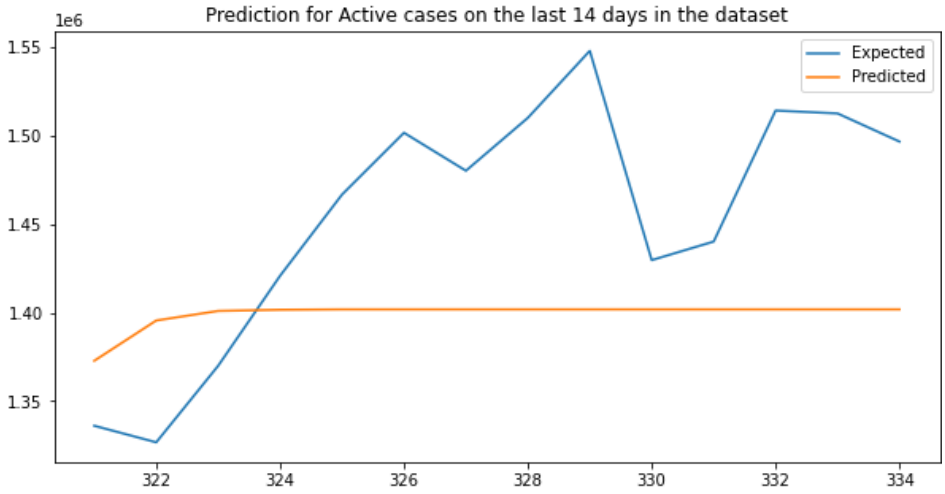


FIGURE 43. Test result of the RFR model for last 14 examples of Brazil data sample.

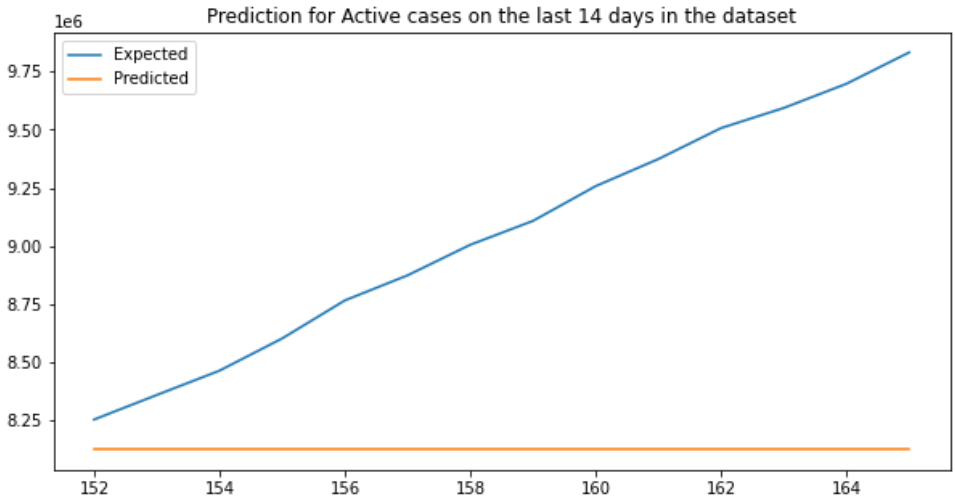


FIGURE 44. Test result of the RFR model for last 14 examples of USA data sample.

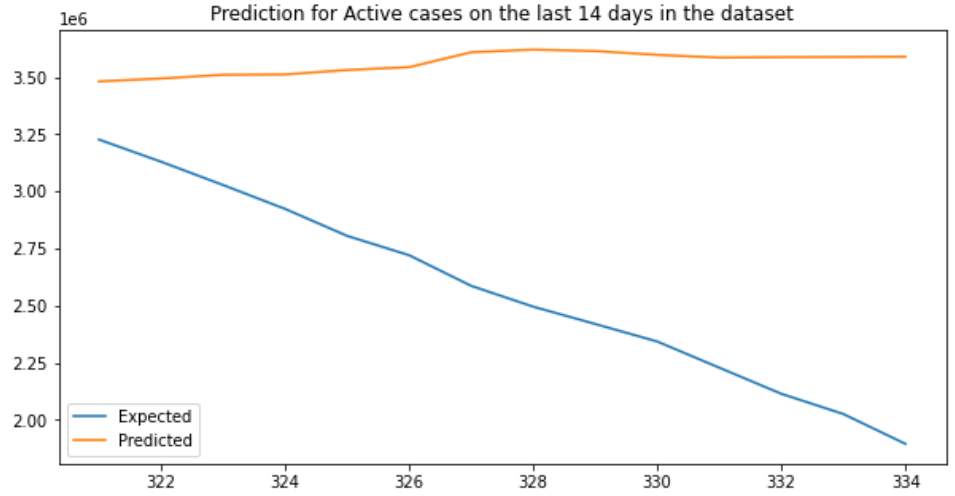


FIGURE 45. Test result of the RFR model for last 14 examples of India data sample.

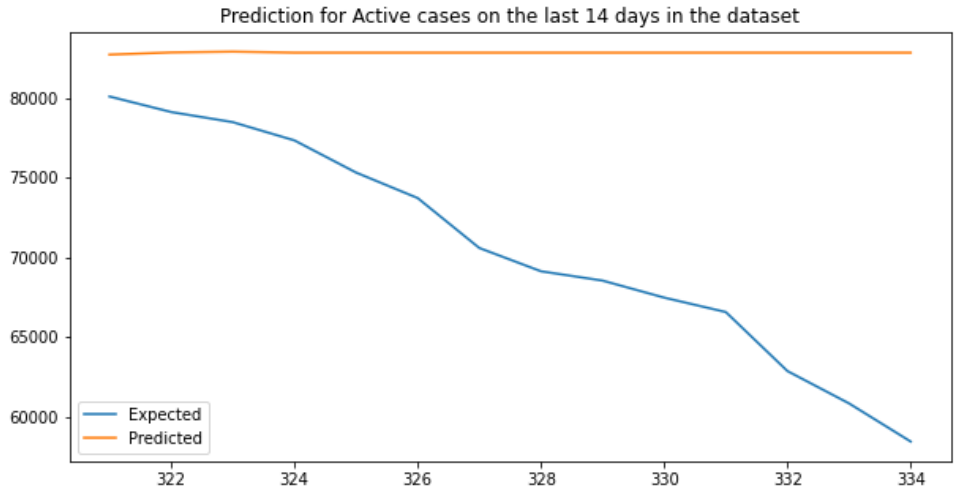


FIGURE 46. Test result of the RFR model for last 14 examples of Japan data sample.

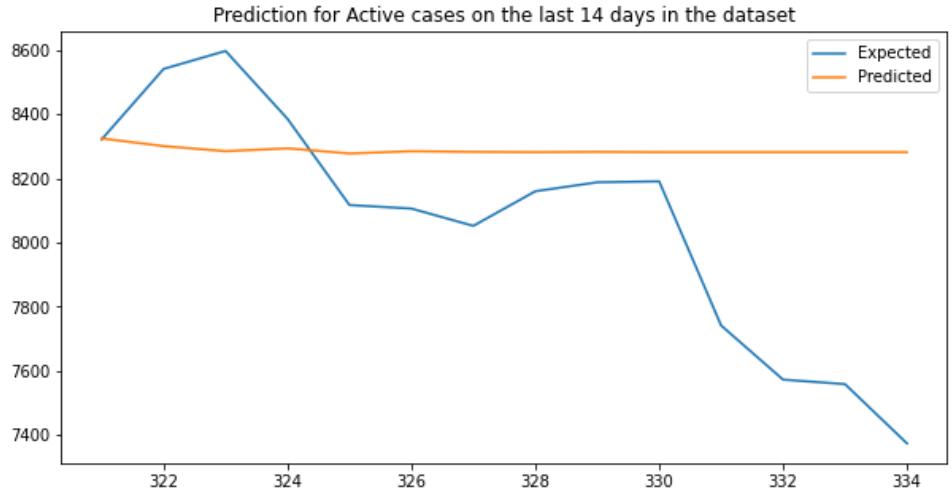


FIGURE 47. Test result of the RFR model for last 14 examples of Korea data sample.

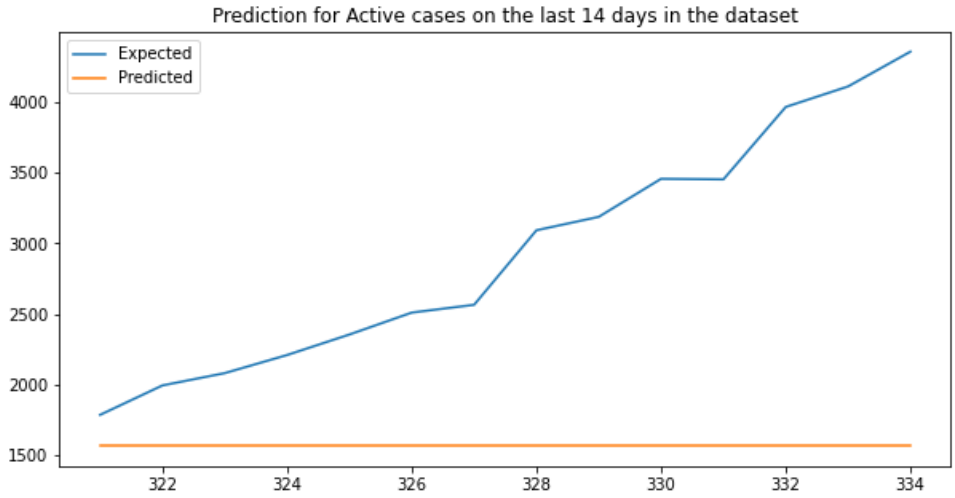


FIGURE 48. Test result of the RFR model for last 14 examples of Vietnam data sample.

The table shows R2 and RMSE evaluation of each cases:

Country	**R^2 Score**	**RMSE Score**	**Range of data**
Global	-8.608	1123005	4585028 - 60675274
Brazil	-0.538	83295.748	344856 - 1548092
United States	-3.474	1050981.423	1823385 - 32673092
India	-6.180	1088871.721	135926 - 3745237
Japan	-3.287	13948.846	1299 - 83832
South Korea	-0.357	418.780	623 - 18073
Vietnam	-2.868	1593.404	15 - 4356

TABLE 2. Score evaluation of RFR for each dataset.

5.1.4 14-day-forward prediction

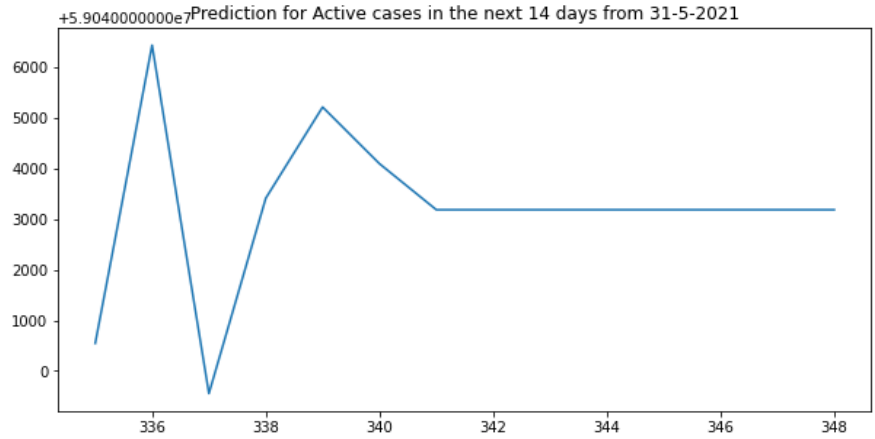


FIGURE 49. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (Global).

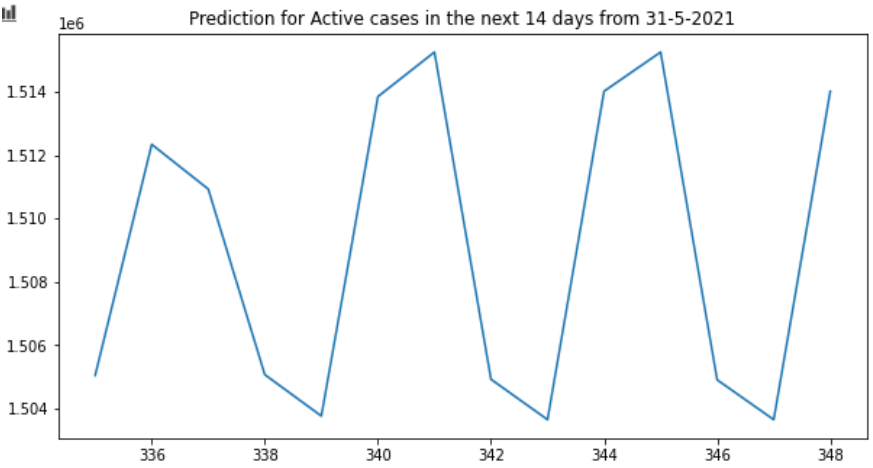


FIGURE 50. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (Brazil).

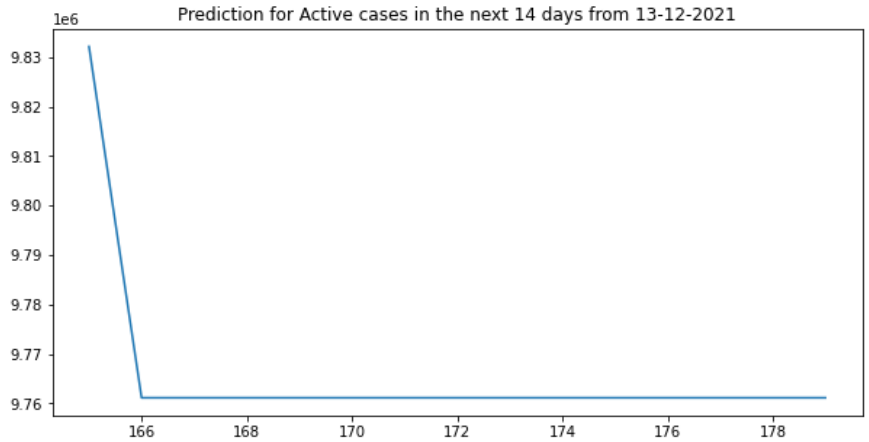


FIGURE 51. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (United States).

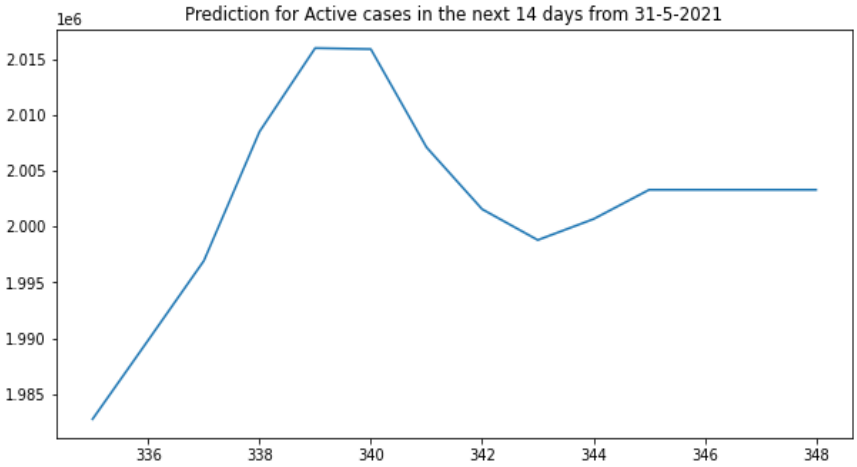


FIGURE 52. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (India).

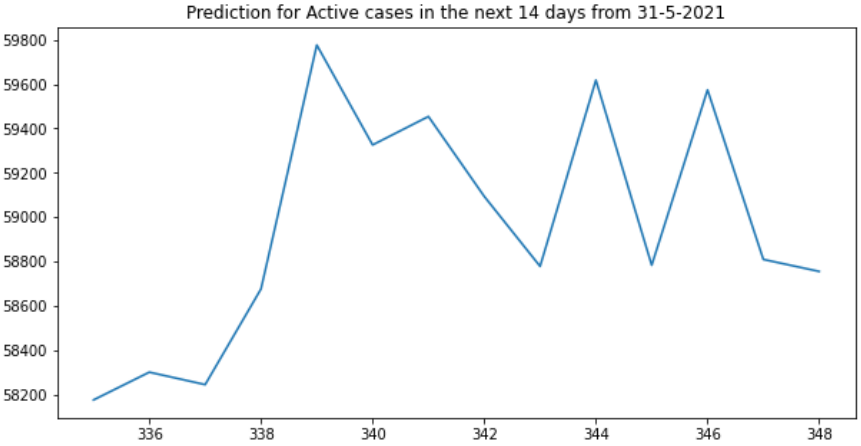


FIGURE 53. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (Japan).

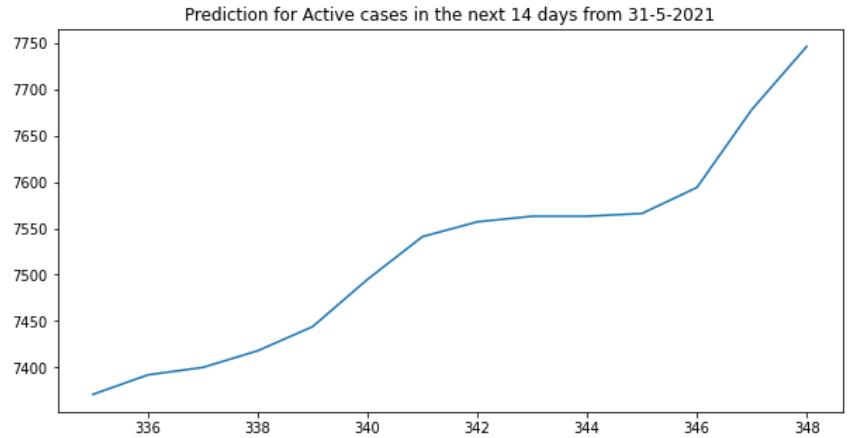


FIGURE 54. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (South Korea).

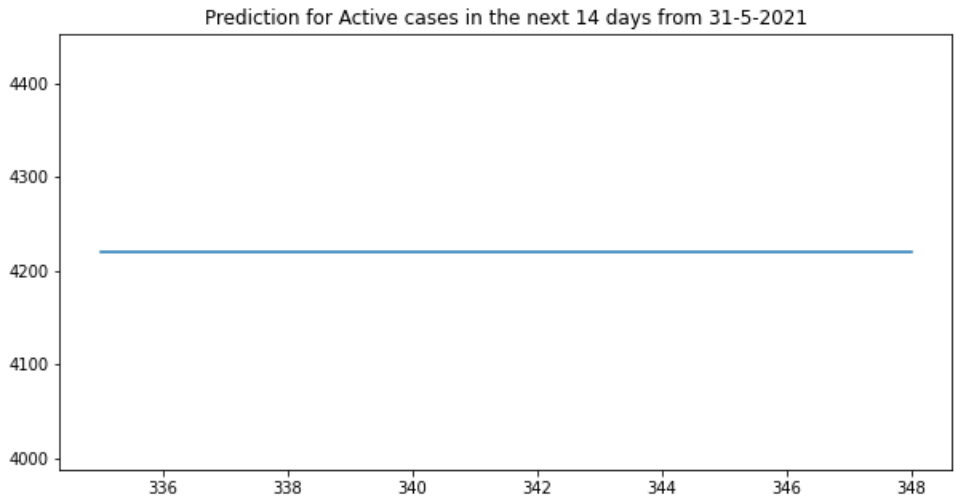


FIGURE 55. Prediction result of the RFR model of data of the next 14 days from May 31, 2021 (Vietnam).

5.1.5 Observations from the model

The model shows negative R^2 results. R^2 scores should varies between 0 and 1. In training progress of all datasets, the predicted value is far from the expected value. In the evaluation, forecast values all seem to run in the opposite direction to the expected value.

In the prediction for the next 14 days after 31-5-2021, the value fluctuates around 3000 more cases then becomes stable. The prediction graphs are very minuscule for a big pandemic like Covid-19.

This suggested that the overall model cannot be applied for this type of data. Random Forest can be good for classification but not as for regression problem as it does not give precise continuous nature prediction. In case of regression, it doesn't predict beyond the range in the training data, and that they may over fit datasets that are particularly noisy.

5.2 ARIMA Model

5.2.1 What is a Time-series?

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

An observed time series can be decomposed into three components: the trend (long term direction), the seasonal (systematic, calendar related movements) and the irregular (unsystematic, short term fluctuations). [57]

In describing these time series, we have used words such as "trend" and "seasonal" which need to be defined more carefully.

(i) Trend

A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as "changing direction," when it might go from an increasing trend to a decreasing trend.

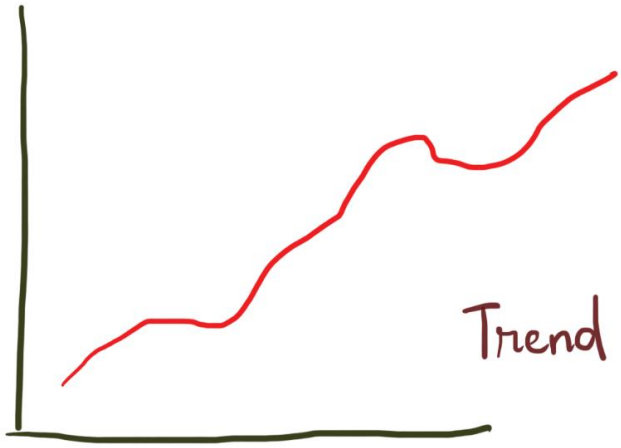


FIGURE 56. Trend pattern of time-series data.

(ii) Seasonal

A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. The monthly sales of antidiabetic drugs above shows seasonality which is induced partly by the change in the cost of the drugs at the end of the calendar year.

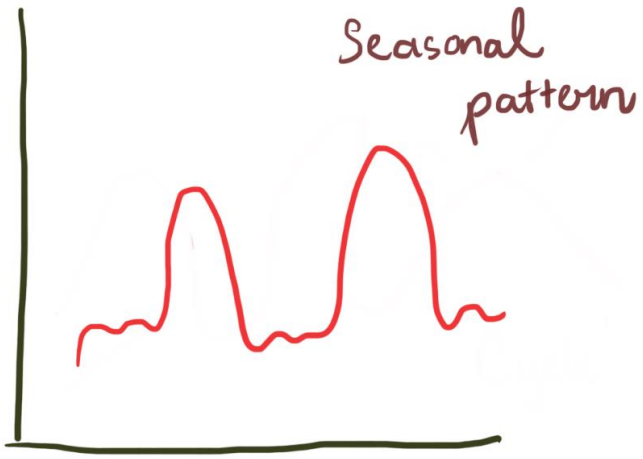


FIGURE 57. Seasonal pattern of time-series data.

(iii) Cyclic

A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to external conditions. Many people confuse cyclic behaviour with seasonal behaviour, but they are really quite different. If the fluctuations are not of a fixed frequency, then they are cyclic; if the frequency is unchanging and associated with some aspect of the calendar, then the pattern is seasonal. In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitudes of cycles tend to be more variable than the magnitudes of seasonal patterns. [58]

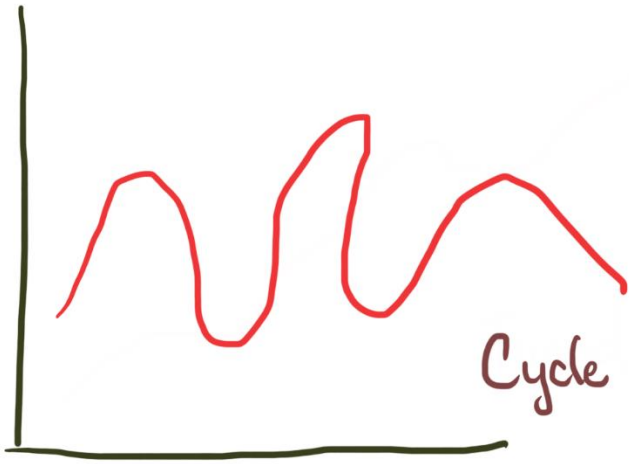


FIGURE 58. Cycle pattern of time-series data.

Time series data is everywhere, since time is a constituent of everything that is observable. As our world gets increasingly instrumented, sensors and systems are constantly emitting a relentless stream of time series data. Such data has numerous applications across various industries. [59]

5.2.2 Data stationarity

A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but in simple language, it is a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations. [60] Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is probably stationary - it should look much the same at any point in time. [61]

In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behaviour is possible), with constant variance.

It is easier to predict when the series is stationary. This is an important step in preparing data to be used in forecasting models such as ARIMA.

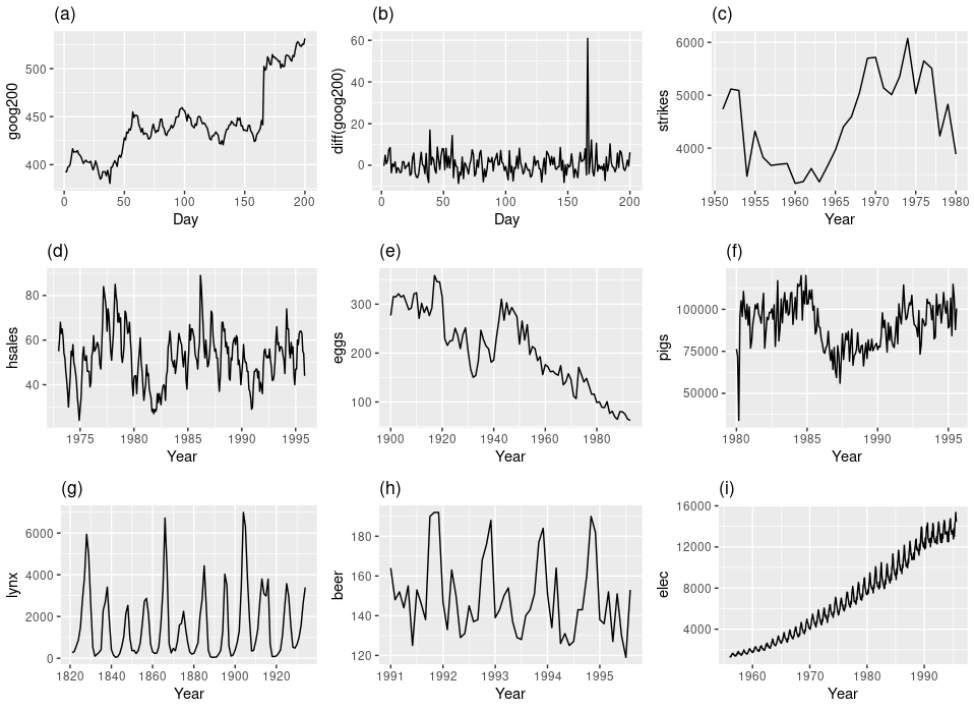


FIGURE 59. Nine time-series plots. Obvious seasonality rules out series (d), (h) and (i). Trends and changing levels rules out series (a), (c), (e), (f) and (i). Increasing variance also rules out (i). That leaves only (b) and (g) as **stationary** series.

5.2.3 ARIMA model definition and formulas

ARIMA, Autoregressive Integrated Moving Average, was introduced by Box and Jenkins in the 1970s. This model takes into consideration changing disturbances in time and tendencies.

ARIMA is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

"MA" is the present value of a series, which is defined as a linear combination of past errors. Assuming the errors to be independently distributed with the normal distribution [13,19], order q is defined as:

$$yt = c + \varepsilon t + \theta 1 yt-1 + \theta 2 yt-2 + \ldots\ldots\ldots\ldots + \theta q yt-q$$

Where:

$$\varepsilon t: \textit{ white noise}$$

$$yt - 1 \textit{ and } yt - 2 \textit{ : lags}$$

Order q of the MA process is obtained from the autocorrelation function (ACF) plot; this is the lag after which ACF crosses the upper confidence interval for the first time. We

combined differencing with MA and AR models, and the combined model can be expressed as:

$$y't = c + \phi 1y't-1 + \phi 2y't-2 + \ldots + \phi py't-p + \theta 1yt-1 + \theta 2yt - 2 + \ldots + \theta qyt-q + \varepsilon t$$

Here, y′t is the differenced series. The "predictors" on the right-hand side include both lagged values of yt and lagged errors. We call this an ARIMA (p, d, q) model, where:

q: order of the MA part

d: degree of first differencing involved

p: order of the AR part

'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors

The value of d, therefore, is the minimum number of differencing needed to make the series stationary.

'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model. [62]

5.2.4 Autocorrelation graph

Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series.

Autocorrelation, as a statistical concept, is also known as serial correlation. The analysis of autocorrelation helps to find repeating periodic patterns, which can be used as a tool of technical analysis in the capital markets.

In many cases, the value of a variable at a point in time is related to the value of it at a previous point in time. Autocorrelation analysis measures the relationship of the observations between the different points in time, and thus seeks for a pattern or trend over the time series. For example, the temperatures on different days in a month are autocorrelated.

Similar to correlation, autocorrelation can be either positive or negative. It ranges from -1 (perfectly negative autocorrelation) to 1 (perfectly positive autocorrelation). Positive autocorrelation means that the increase observed in a time interval leads to a proportionate increase in the lagged time interval. [63]

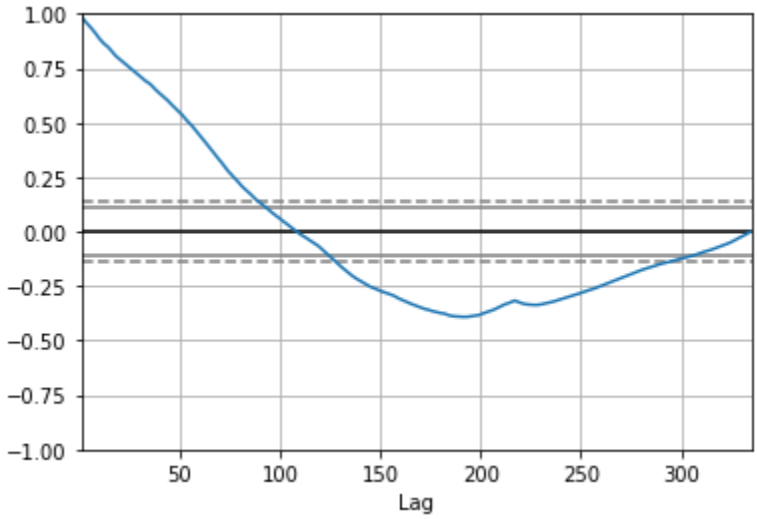


FIGURE 60: Autocorrelation plot.

5.2.5 Pre-proposed method

Some steps shall be performed before defining the best fit for the model.

5.2.5.1 Stationarity verification

First of all, time-series stationarity and seasonality need to be verified. To determine whether a given series is stationary or not and deal with it accordingly, some methods are introduced. For simplicity, we only focus on one main method: the **ADF (Augmented Dickey Fuller)** test.

It is one of the most popular statistical tests. It can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The null and alternate hypothesis of this test are:

_ *Null Hypothesis: The series has a unit root (value of a =1)*

_ *Alternate Hypothesis: The series has no unit root.*

A result of ADF test run is simulate below:

```
Test Statistic                  0.815369
p-value                         0.991880
#Lags Used                     13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
```

FIGURE 61. A sample output after ADF test run. We mainly consider the Test statistic and p-value.

Basically, if the test statistic is less than the critical value, we can reject the null hypothesis (aka the series is stationary). When the test statistic is greater than the critical value, we fail to reject the null hypothesis (which means the series is not stationary).

If we fail to reject the null hypothesis, we can say that the series is non-stationary. [64]

5.2.5.2 Data transformation methods

The two most common ways to make a non-stationary time series curve stationary are:

(i) Differencing

In order to make series stationary, we take a difference between the data points. For example, the original time series is:

$$x_1, x_2, x_3, \ldots, x_n$$

The series with difference of degree 1 becomes:

$$x_2 - x_1, x_3 - x_2, x_4 - x_3, \ldots, x_n - x_{n-1}$$

Once, take the difference, plot the series and see if there is any improvement in the curve. If not, second or even third order differentiation is normally used. However, the more differences, the more complicated the analysis becomes.

(ii) Transforming

In case differencing is not working, it is usually suggested that we transform the variables. **Logarithm transform** is probably the most commonly used transformation, along with basic reciprocal, square root, cube root, or power function methods.

One of the more complex and commonly used formulas is the Box-Cox transformation, named after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed the technique. [65] Let w be the transformed variable and y is the target variable, then:

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^\lambda - 1)/\lambda & \text{otherwise.} \end{cases}$$

FIGURE 62. Box-Cox Transformation mathematical expression. Photo from Rob Hyndman's and George Athanasopoulos's "Forecasting".

where t is the time period and lambda is the customized parameter that we have to choose. We choose the value of lambda that provides the best approximation for the normal distribution of our response variable. [66] In Python, this can be done automatically by applying pre-defined function *scipy.stats.boxcox()* [67] from the *scipy* library.

Data used for ARIMA model in this study were transformed using Yeo-Johnson transformation. This method is quite similar to the Box-Cox transformation, but it handles

both positive and negative values, whereas the Box-Cox transformation only handles positive values (different from 0).

5.2.5.3 Estimation method for p, d, q and evaluation criteria

The approximation technique considered for defining the three parameters (p, q, and d) of the ARIMA model was **Grid Search CV**, which tests a wide set of parameter-matching implementations and selects the best fit based on the lowest Root Mean-Squared Error (RMSE) score.

The evaluation criteria also taking into account the Mean Absolute Error (MAE) and R-squared Error (R^2).

5.2.5.4 Train-test split method

In applied machine learning, we often split our data into a train and a test set: the training set used to prepare the model and the test set used to evaluate it. We may even use k-fold cross validation that repeats this process by systematically splitting the data into k groups, each of which has a chance of being a sustained model. However, these methods cannot be used directly with time series data. We cannot select random samples and assign them neither to the test set nor to the train set, since it makes no sense to use the values from the future to predict values in the past. In short, we don't want any future prospects when training our model. [68]

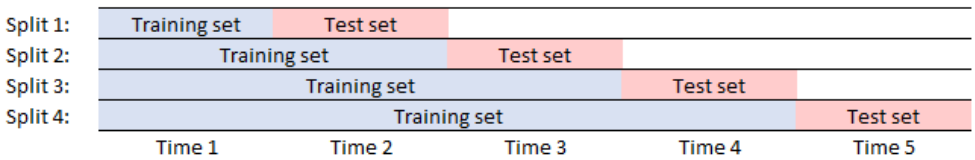


FIGURE 63. Roll-forward train-test validation, visually explained.

Instead, we use cross-validation on a rolling-forward basis. Start with a small subset of data for training purpose, forecast the later data points, and then verify the accuracy of the forecasted data points. The same predicted data points are then included as part of the next training dataset and subsequent data points are forecasted.

All datasets are split into 5 smaller train-test samples, with test size always equals 55 values. The numbers indicate indexes:

+ *Sample 1: train [0-60], test [60-115]*
+ *Sample 2: train [0-115], test [115-170]*
+ *Sample 3: train [0-170], test [170-225]*
+ *Sample 4: train [0-225], test [225-280]*
+ *Sample 5: train [0-280], test [280-335].*

5.2.6 Results

The results were organized in the following sequence: Brazil, United States, India, Japan, South Korea, Vietnam.

5.2.6.1 Brazil

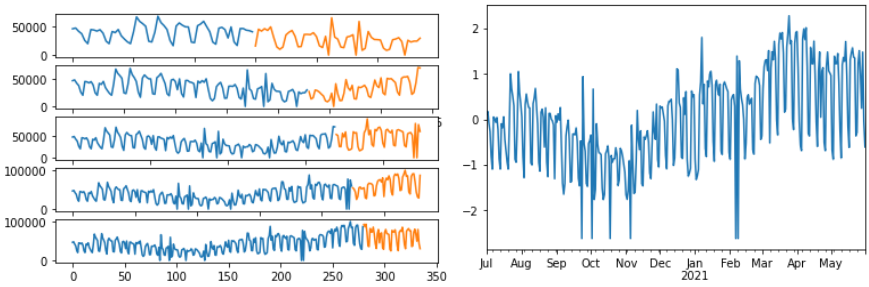


FIGURE 64. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

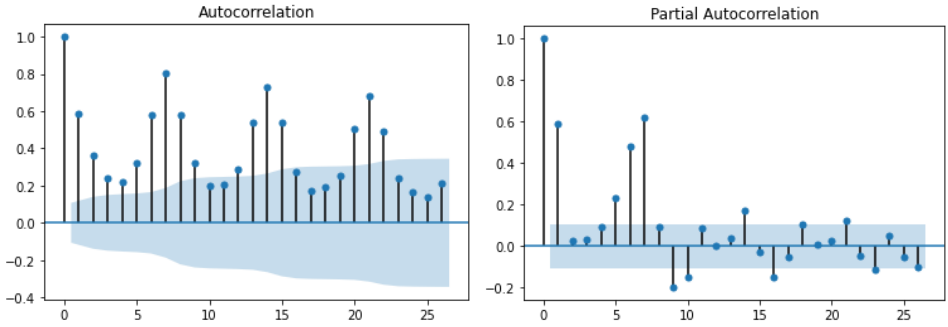


FIGURE 65. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

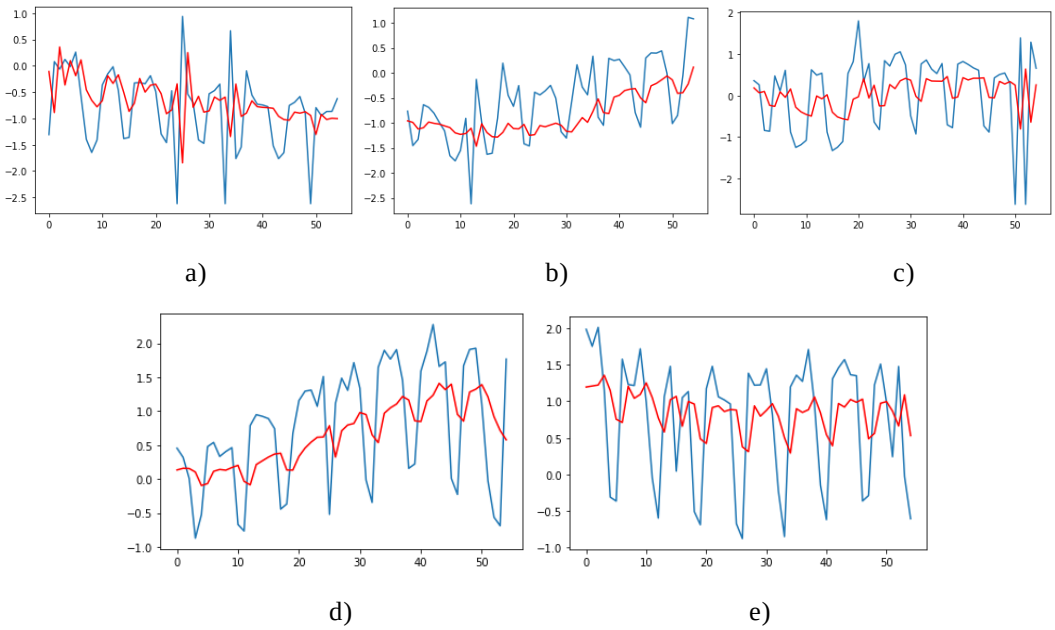


FIGURE 66. Prediction results of the ARIMA (p = 0, q = 1, d = 2) model for five sample of COVID-19 new cases time series from Brazil. The blue line is the original data, while the red line is the prediction time series.

Brazil	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.96	0.758	1.074	1.343	1.195
ARIMA (0,0,1)	0.993	0.678	1.068	1.096	0.955
ARIMA (0,0,2)		0.672	1.017	1.008	0.951
ARIMA (0,0,3)	1.02	0.668	1.008	0.918	0.909
ARIMA (0,1,0)	1.183	0.723	1.293	0.87	0.973
ARIMA (0,1,1)	0.949	0.69	0.988	0.844	0.886
ARIMA (0,1,2)	**0.921**	**0.67**	**1.002**	**0.795**	**0.819**
ARIMA (0,1,3)	0.935		1.001	0.797	0.836
ARIMA (0,2,0)	2.074	1.131	2.248	1.18	1.429
ARIMA (0,2,1)	1.197	0.728	1.3	0.875	0.978
ARIMA (1,0,0)	0.955	0.663	1.041	0.959	0.897
ARIMA (1,0,1)			1.009	0.861	0.883
ARIMA (1,0,2)			1.001	0.814	0.815
ARIMA (1,0,3)			1.001	0.817	0.831
ARIMA (1,1,0)	1.085	0.714	1.137	0.928	1.021
ARIMA (1,1,1)					
ARIMA (1,1,2)			1.006	0.794	0.823
ARIMA (1,1,3)			1.032	0.795	0.765
ARIMA (1,2,0)	1.553	0.99	1.481	1.236	1.427
ARIMA (2,0,0)	1.007	0.665	1.033	0.944	0.921
ARIMA (2,0,1)			0.993	0.816	0.831
ARIMA (2,0,2)			0.947	0.925	0.818
ARIMA (2,0,3)					0.825
ARIMA (2,1,0)	1.097	0.69	1.162	0.92	0.981
ARIMA (2,1,1)					
ARIMA (2,1,2)			1.004	0.702	0.714
ARIMA (2,1,3)					
ARIMA (2,2,0)	1.433	0.885	1.451	1.181	1.297
ARIMA (2,2,1)					
ARIMA (2,2,2)					

TABLE 3: Best fit of ARIMA orders for each sample (orange cells). At order (0,1,2), the differences (in RMSE) between chosen values (same row as (0,1,2)) and best fit values are relatively small. Grey cells indicate no result.

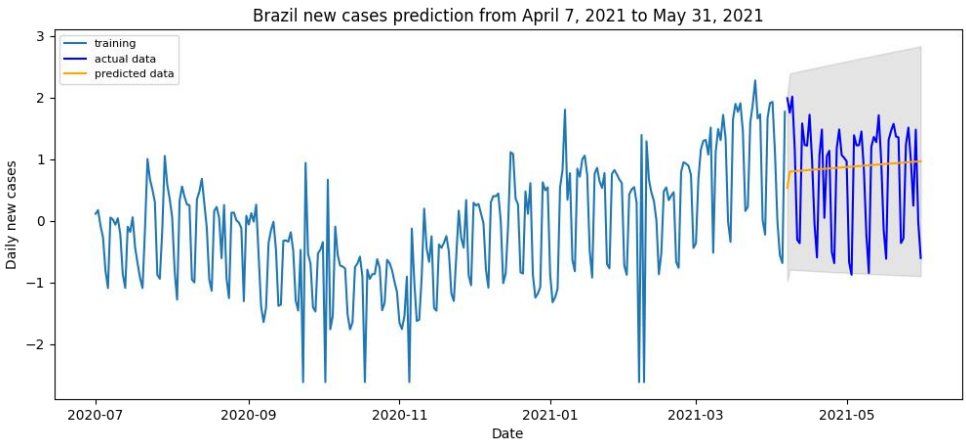


FIGURE 67. Prediction compared to actual data on the last 55 days of dataset.

5.2.6.2 United States

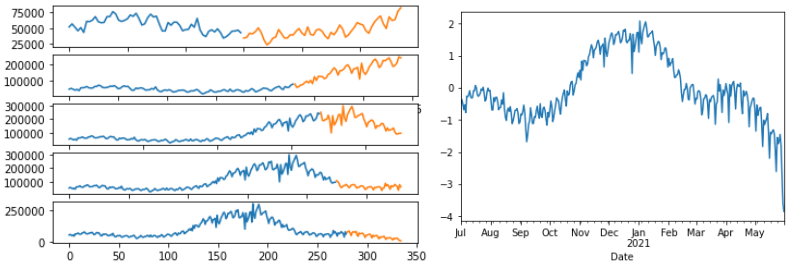


FIGURE 68. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

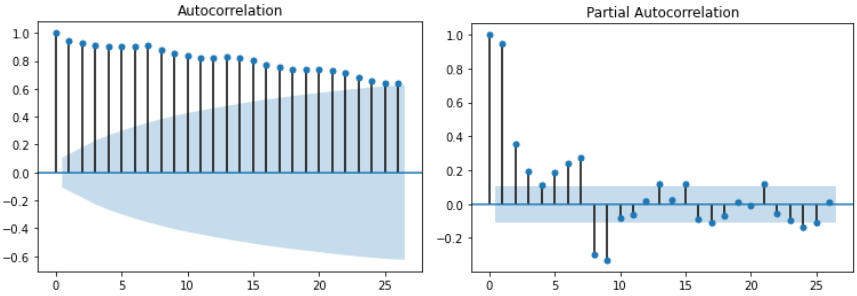


FIGURE 69. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

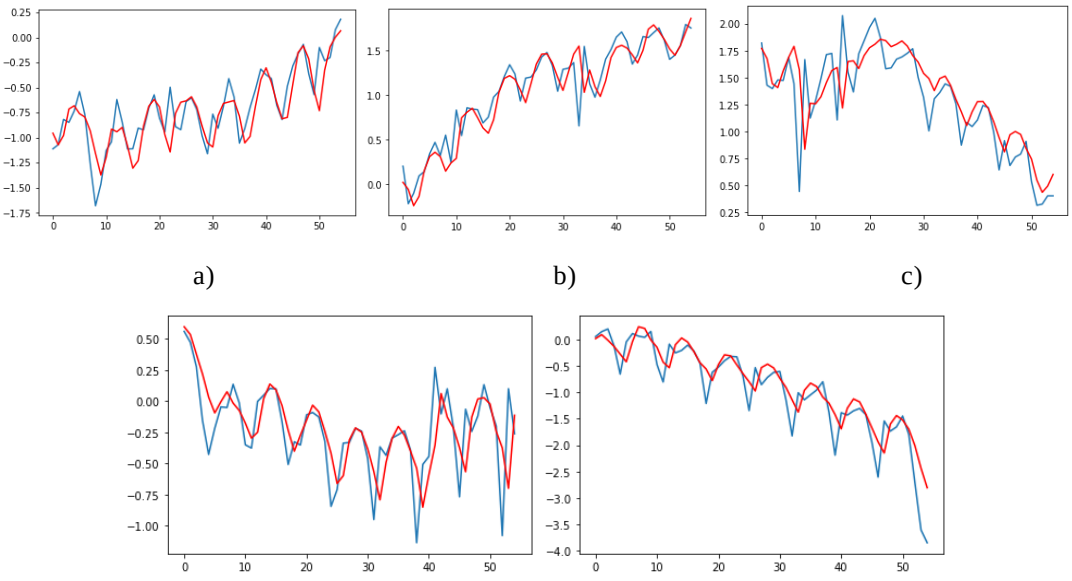


FIGURE 70. Prediction results of the ARIMA (p = 2, q = 1, d = 3) model for five sample of COVID-19 new cases time series from the United States. The blue line is the original data, while the red line is the prediction time series.

US	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.42	1.45	1.234	0.576	1.377
ARIMA (0,0,1)	0.288	0.87	0.754	0.458	0.858
ARIMA (0,0,2)			0.599	0.386	0.657
ARIMA (0,0,3)			0.561	0.389	0.613
ARIMA (0,1,0)	0.247	0.248	0.336	0.361	0.452
ARIMA (0,1,1)	0.25	0.238	0.304	0.334	0.472
ARIMA (0,1,2)	0.227	0.239	0.303	0.325	0.451
ARIMA (0,1,3)	0.23	0.239	0.308	0.322	0.449
ARIMA (0,2,0)	0.339	0.421	0.57	0.583	0.668
ARIMA (0,2,1)	0.251	0.25	0.339	0.363	0.45
ARIMA (1,0,0)	0.24	0.269	0.331	0.354	0.459
ARIMA (1,0,1)	0.239	0.266	0.301	0.332	0.477
ARIMA (1,0,2)	0.233			0.323	0.456
ARIMA (1,0,3)	0.229			0.32	0.454
ARIMA (1,1,0)	0.249	0.242	0.314	0.345	0.458
ARIMA (1,1,1)		0.229	0.302	0.324	0.451
ARIMA (1,1,2)		0.24	0.306	0.324	0.45
ARIMA (1,1,3)		0.244	0.31	0.322	0.449
ARIMA (1,2,0)	0.329	0.326	0.441	0.45	0.588
ARIMA (2,0,0)	0.24	0.267	0.311	0.341	0.465
ARIMA (2,0,1)		0.244		0.322	0.456
ARIMA (2,0,2)		0.251	0.3	0.322	0.455
ARIMA (2,0,3)		0.258		0.32	0.458
ARIMA (2,1,0)	0.241	0.247	0.311	0.345	0.457
ARIMA (2,1,1)		0.243	0.308	0.324	0.449
ARIMA (2,1,2)		0.229	0.3	0.302	0.42
ARIMA (2,1,3)	**0.205**	**0.205**	**0.288**	**0.263**	**0.373**
ARIMA (2,2,0)	0.302	0.317	0.391	0.437	0.538
ARIMA (2,2,1)					
ARIMA (2,2,2)					

TABLE 4: Best fit of ARIMA orders for each sample (green cells). At order (2,1,3), the differences (in RMSE) between chosen values (same row as (2,1,3)) and best fit values are relatively small. Grey cells indicate no result.

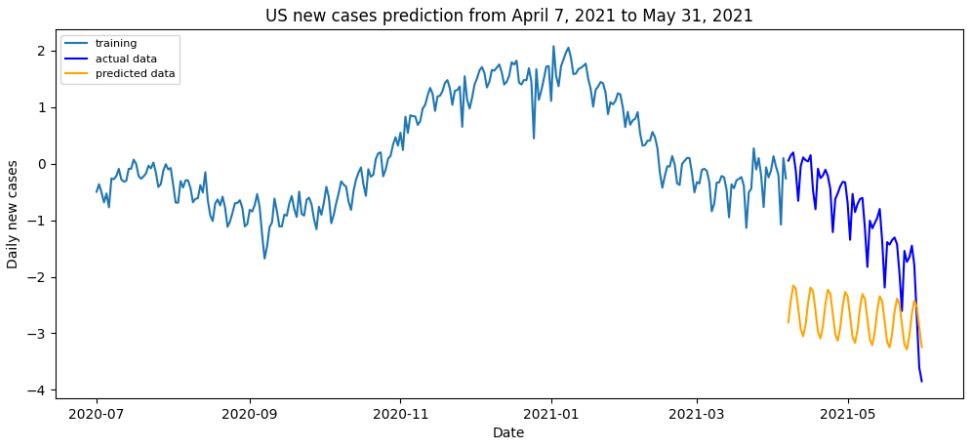


FIGURE 71. Prediction compared to actual data on the last 55 days of dataset.

5.2.6.1 India

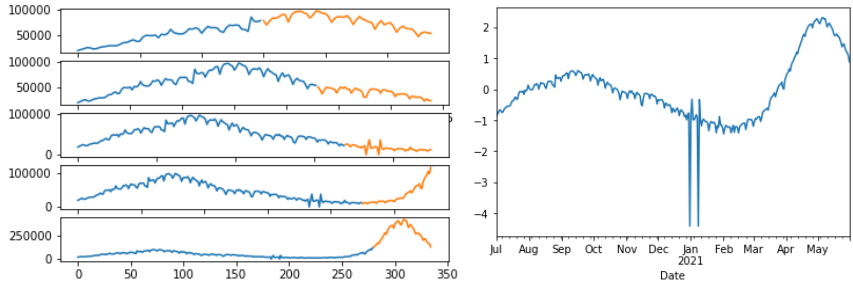


FIGURE 72. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

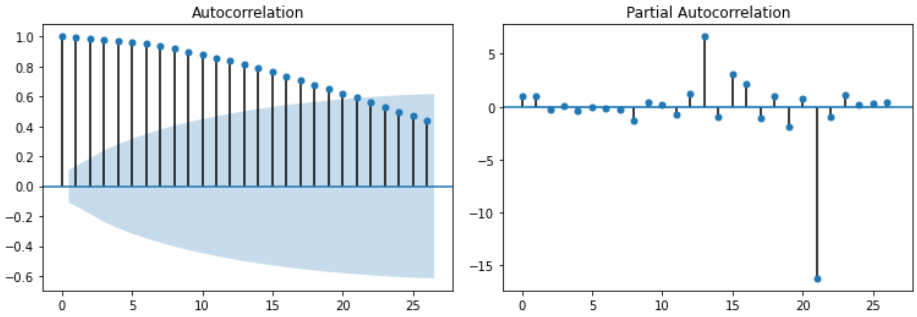


FIGURE 73. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

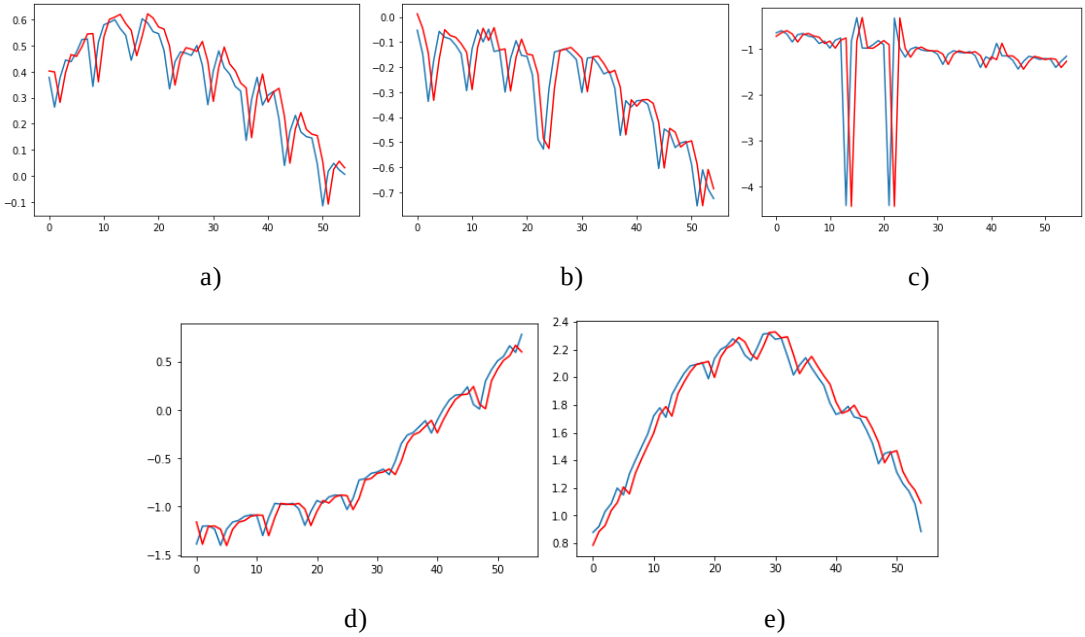


FIGURE 74. Prediction results of the ARIMA (p = 0, q = 1, d = 0) model for five sample of COVID-19 new cases time series from India. The blue line is the original data, while the red line is the prediction time series.

India	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.391	0.376	1.18	0.663	1.963
ARIMA (0,0,1)	0.231	0.214	1.09	0.447	1.217
ARIMA (0,0,2)	0.162	0.163	1.188	0.365	0.902
ARIMA (0,0,3)				0.342	0.8
ARIMA (0,1,0)	**0.091**	**0.105**	**1.022**	**0.115**	**0.09**
ARIMA (0,1,1)	0.093	0.107	0.772	0.215	0.227
ARIMA (0,1,2)	0.089	0.102			
ARIMA (0,1,3)	0.09	0.103			
ARIMA (0,2,0)	0.133	0.152	1.77	0.164	0.098
ARIMA (0,2,1)		0.105	1.028	0.117	0.091
ARIMA (1,0,0)	0.088	0.104	0.999	0.207	0.302
ARIMA (1,0,1)	0.088	0.105		0.218	0.234
ARIMA (1,0,2)	0.182	0.1		0.222	0.233
ARIMA (1,0,3)	0.15	0.101		0.228	0.231
ARIMA (1,1,0)	0.092	0.106	1.043	0.125	0.114
ARIMA (1,1,1)	0.092	0.104		0.217	0.227
ARIMA (1,1,2)	0.09	0.103		0.187	0.17
ARIMA (1,1,3)	0.082	0.103		0.188	0.174
ARIMA (1,2,0)	0.124	0.138	1.573	0.151	0.096
ARIMA (2,0,0)	0.088	0.105	1.018	0.161	0.198
ARIMA (2,0,1)	2.245	0.102		0.22	24059.45
ARIMA (2,0,2)	0.084	0.1			0.167
ARIMA (2,0,3)		0.103		0.193	0.171
ARIMA (2,1,0)	0.087	0.104	0.97	0.139	0.141
ARIMA (2,1,1)	0.09	0.103		0.218	0.225
ARIMA (2,1,2)		0.097		0.188	0.173
ARIMA (2,1,3)		0.104			0.171
ARIMA (2,2,0)	0.104	0.128		0.138	0.094
ARIMA (2,2,1)					
ARIMA (2,2,2)		0.103			

TABLE 5: Best fit of ARIMA orders for each sample (pink cells). At order (0,1,0), the differences (in RMSE) between chosen values (same row as (0,1,0)) and best fit values are relatively small. Grey cells indicate no result.

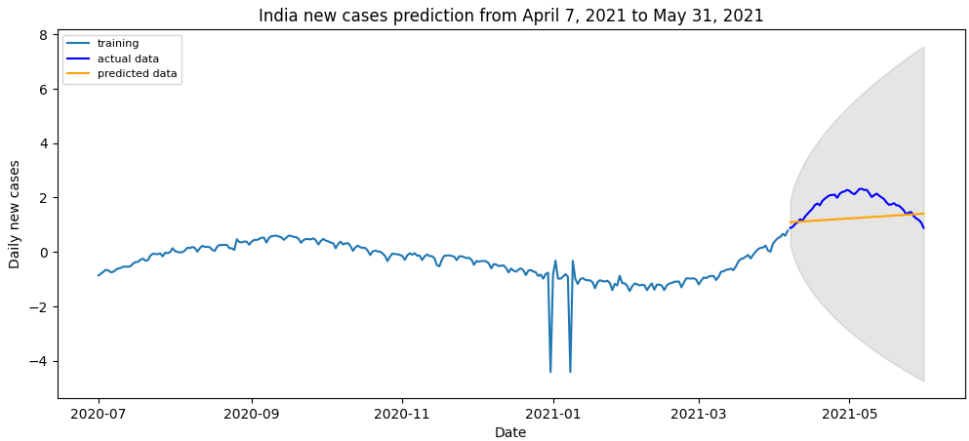


FIGURE 75. Prediction compared to actual data on the last 55 days of dataset.

5.2.6.1 Japan

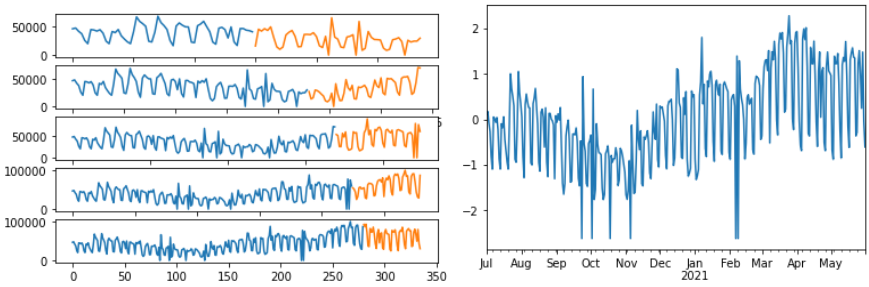


FIGURE 76. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

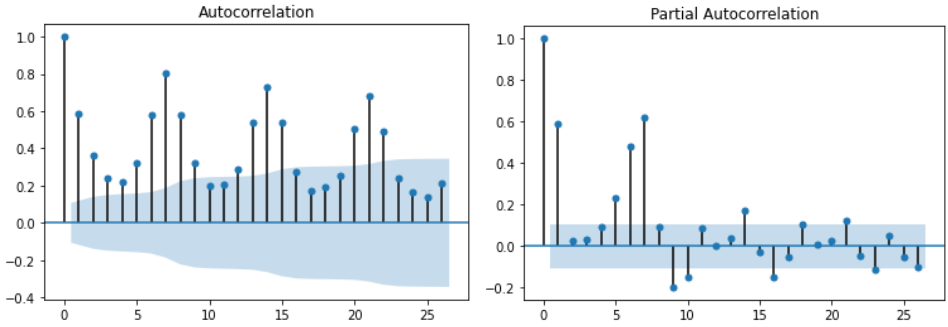


FIGURE 77. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

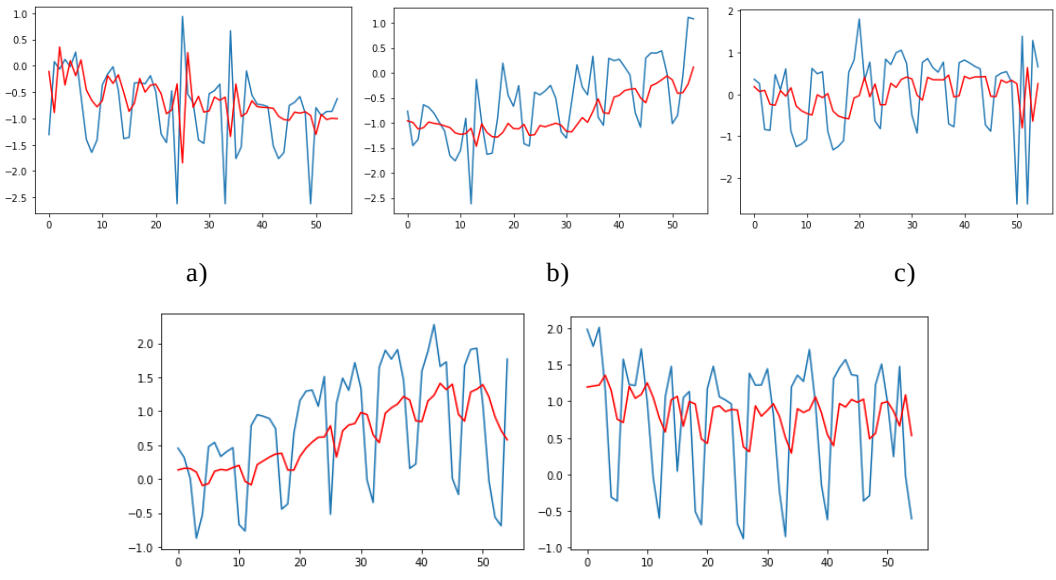


FIGURE 78. Prediction results of the ARIMA (p = 2, q = 1, d = 3) model for five sample of COVID-19 new cases time series from Japan. The blue line is the original data, while the red line is the prediction time series.

Japan	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.37	1.016	1.531	0.438	1.392
ARIMA (0,0,1)	0.302	0.621	0.843	0.566	0.834
ARIMA (0,0,2)	0.291	0.492	0.607	0.445	0.596
ARIMA (0,0,3)	0.286	0.439			0.559
ARIMA (0,1,0)	0.324	0.286	0.234	0.4	0.331
ARIMA (0,1,1)	0.326	0.293	0.264	0.353	0.319
ARIMA (0,1,2)	0.304	0.263	0.252	0.36	0.308
ARIMA (0,1,3)	0.306	0.263	0.249	0.36	0.308
ARIMA (0,2,0)	0.481	0.407	0.301	0.685	0.515
ARIMA (0,2,1)	0.331	0.29	0.235	0.402	0.332
ARIMA (1,0,0)	0.301	0.311	0.245	0.391	0.332
ARIMA (1,0,1)	0.317	0.313	0.253	0.357	0.321
ARIMA (1,0,2)	0.295	0.28	0.244	0.36	0.309
ARIMA (1,0,3)		0.281	0.241	0.359	0.309
ARIMA (1,1,0)	0.328	0.292	0.241	0.375	0.325
ARIMA (1,1,1)	0.307	0.269	0.253	0.35	0.309
ARIMA (1,1,2)	0.307	0.263	0.25	0.36	0.308
ARIMA (1,1,3)	0.3	0.252	0.264	0.36	0.307
ARIMA (1,2,0)	0.426	0.384	0.289	0.517	0.432
ARIMA (2,0,0)	0.308	0.311	0.247	0.372	0.326
ARIMA (2,0,1)	0.294	0.286	0.245	0.35	0.31
ARIMA (2,0,2)	0.296	0.28	0.242	0.36	0.309
ARIMA (2,0,3)		0.265	0.24	0.359	0.308
ARIMA (2,1,0)	0.327	0.284	0.244	0.38	0.32
ARIMA (2,1,1)	0.303	0.258	0.249	0.363	0.307
ARIMA (2,1,2)	0.314	0.238	0.251	0.347	0.288
ARIMA (2,1,3)	**0.318**	**0.242**	**0.214**	**0.313**	**0.286**
ARIMA (2,2,0)	0.411	0.362	0.291	0.49	0.402
ARIMA (2,2,1)					
ARIMA (2,2,2)					

TABLE 6: Best fit of ARIMA orders for each sample (blue cells). At order (2,1,3), the differences (in RMSE) between chosen values (same row as (2,1,3)) and best fit values are relatively small. Grey cells indicate no result.

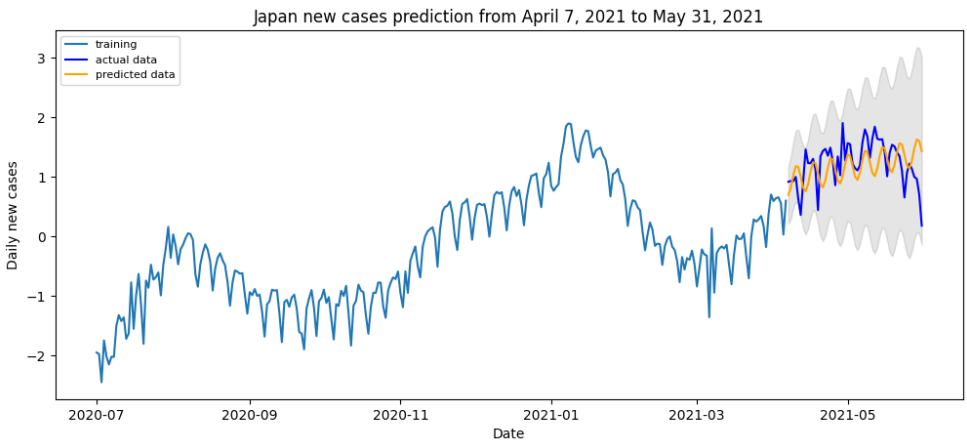


FIGURE 79. Prediction compared to actual data on the last 55 days of dataset.

5.2.6.1 South Korea

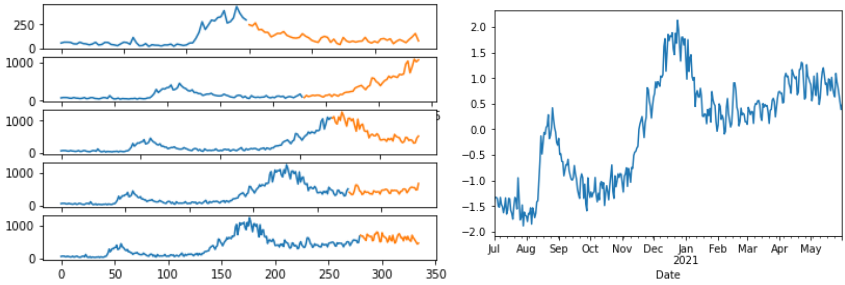


FIGURE 80. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

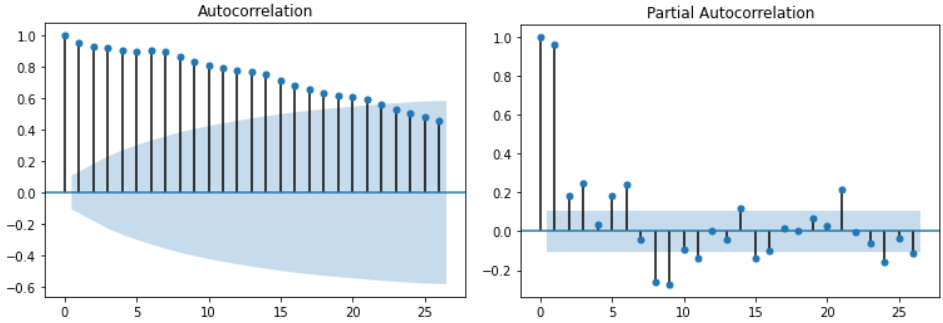


FIGURE 81. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

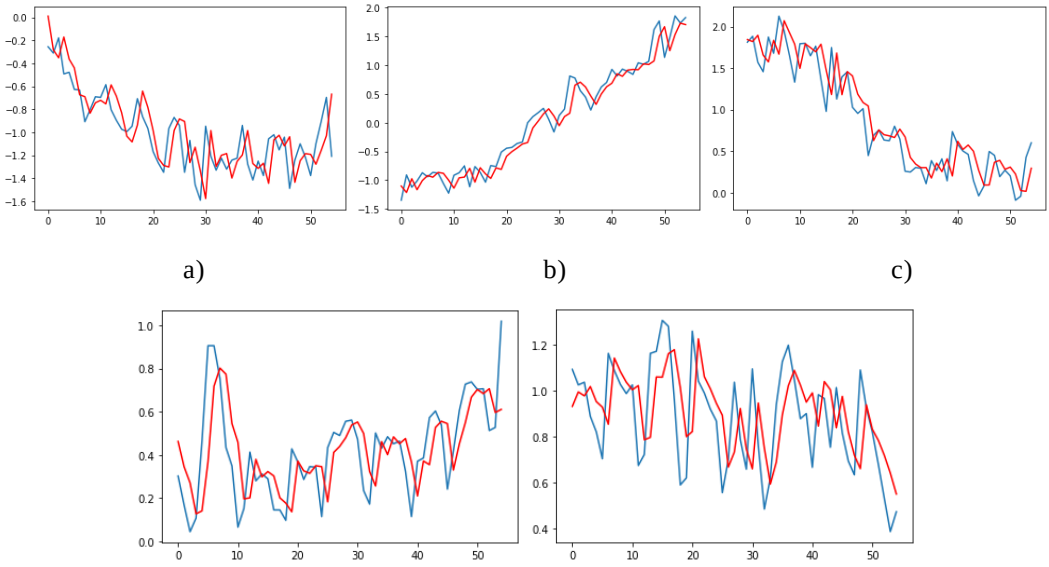


FIGURE 82. Prediction results of the ARIMA (p = 2, q = 1, d = 2) model for five sample of COVID-19 new cases time series from South Korea. The blue line is the original data, while the red line is the prediction time series.

Korea	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.333	1.327	1.505	0.699	0.995
ARIMA (0,0,1)	0.253	0.777	0.874	0.412	0.571
ARIMA (0,0,2)	0.266	0.581	0.642	0.305	0.408
ARIMA (0,0,3)	0.246			0.268	
ARIMA (0,1,0)	0.226	0.228	0.288	0.185	0.229
ARIMA (0,1,1)	0.216	0.225	0.272	0.192	0.227
ARIMA (0,1,2)	0.216	0.221	0.272	0.182	0.213
ARIMA (0,1,3)	0.217	0.221	0.272	0.185	0.214
ARIMA (0,2,0)	0.349	0.352	0.466	0.243	0.345
ARIMA (0,2,1)	0.23	0.228	0.292	0.186	0.23
ARIMA (1,0,0)	0.214	0.245	0.284	0.185	0.227
ARIMA (1,0,1)	0.208			0.192	0.225
ARIMA (1,0,2)				0.183	0.212
ARIMA (1,0,3)				0.186	0.213
ARIMA (1,1,0)	0.219	0.228	0.277	0.192	0.23
ARIMA (1,1,1)	0.216		0.271	0.184	0.217
ARIMA (1,1,2)		0.22	0.271	0.184	0.214
ARIMA (1,1,3)					0.209
ARIMA (1,2,0)	0.297	0.314	0.379	0.242	0.319
ARIMA (2,0,0)	0.21	0.243	0.272	0.191	0.228
ARIMA (2,0,1)	0.207			0.185	0.216
ARIMA (2,0,2)	0.208			0.185	0.213
ARIMA (2,0,3)	0.21				0.208
ARIMA (2,1,0)	0.216	0.221	0.272	0.19	0.22
ARIMA (2,1,1)	0.217	0.222	0.273	0.186	0.214
ARIMA (2,1,2)	**0.227**	**0.219**	**0.272**	**0.184**	**0.207**
ARIMA (2,1,3)					
ARIMA (2,2,0)	0.266	0.268	0.328	0.244	0.284
ARIMA (2,2,1)	0.224	0.218	0.281		
ARIMA (2,2,2)					

TABLE 7: Best fit of ARIMA orders for each sample (yellow cells). At order (2,1,2), the differences (in RMSE) between chosen values (same row as (2,1,2)) and best fit values are relatively small. Grey cells indicate no result.

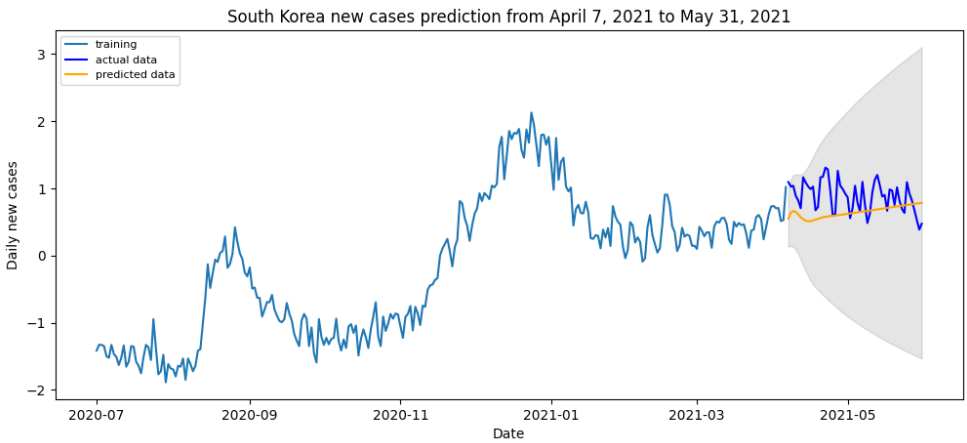


FIGURE 83. Prediction compared to actual data on the last 55 days of dataset.

5.2.6.1 Vietnam

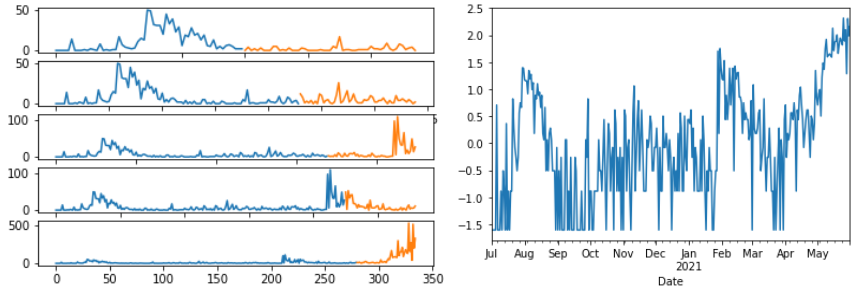


FIGURE 84. Train-test samples visualized (left) and data after transformed using Yeo-Johnson method (right).

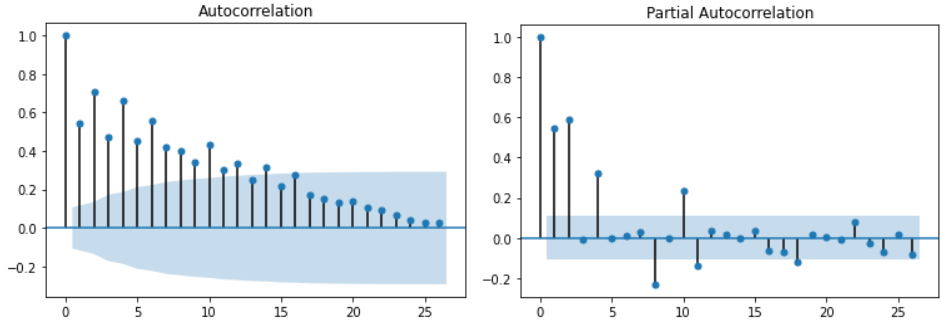


FIGURE 85. Autocorrelation plot (ACF, left) and Partial autocorrelation plot (PACF, right).

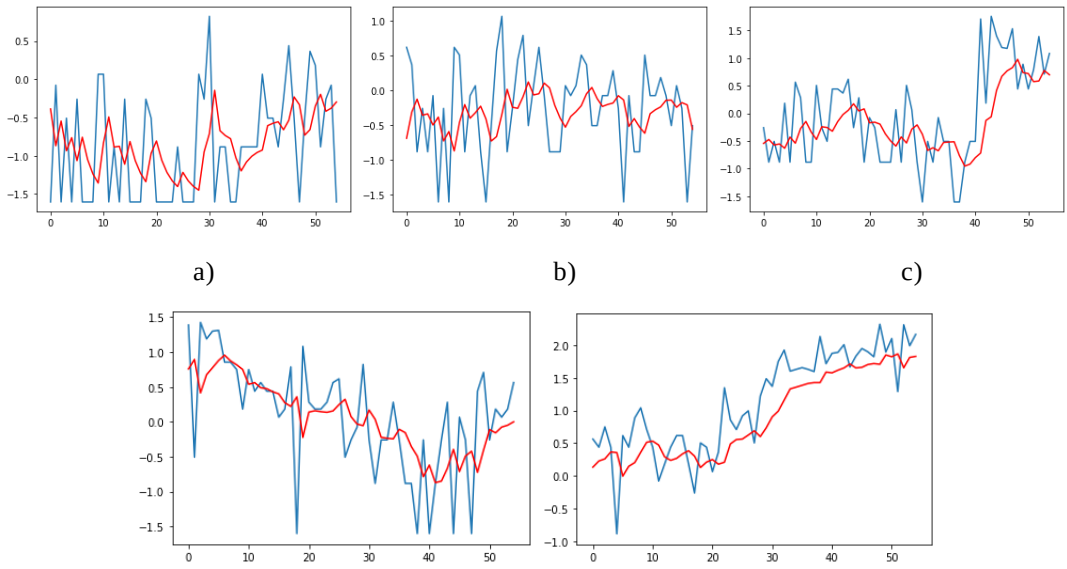


FIGURE 86. Prediction results of the ARIMA (p = 1, q = 0, d = 1) model for five sample of COVID-19 new cases time series from Vietnam. The blue line is the original data, while the red line is the prediction time series.

Vietnam	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
ARIMA (0,0,0)	0.935	0.674	0.91	0.857	1.453
ARIMA (0,0,1)	0.872	0.694	0.814	0.81	1.087
ARIMA (0,0,2)	0.833	0.765	0.756	0.802	0.916
ARIMA (0,0,3)	0.874	0.763	0.731	0.755	0.833
ARIMA (0,1,0)	0.94	0.928	0.807	0.926	0.46
ARIMA (0,1,1)	0.768	0.724	0.7	0.686	0.397
ARIMA (0,1,2)	0.775	0.727	0.701	0.686	0.398
ARIMA (0,1,3)					
ARIMA (0,2,0)	0.586	1.577	1.408	1.631	0.771
ARIMA (0,2,1)	0.953	0.937	0.811	0.929	0.462
ARIMA (1,0,0)	0.832	0.756	0.735	0.784	0.699
ARIMA (1,0,1)	**0.756**	**0.715**	**0.7**	**0.674**	**0.454**
ARIMA (1,0,2)	0.764	0.719	0.701	0.672	0.455
ARIMA (1,0,3)			0.704	0.673	0.456
ARIMA (1,1,0)	0.84	0.846	0.699	0.777	0.425
ARIMA (1,1,1)	0.78	0.728	0.7	0.687	0.398
ARIMA (1,1,2)		0.735	0.7	0.687	0.398
ARIMA (1,1,3)	0.767	0.749	0.697	0.697	0.398
ARIMA (1,2,0)	0.171	1.279	0.995	1.214	0.608
ARIMA (2,0,0)	0.799	0.771	0.682	0.728	0.553
ARIMA (2,0,1)	0.777	0.719	0.701	0.673	0.455
ARIMA (2,0,2)			0.699	0.676	0.463
ARIMA (2,0,3)			0.699	0.682	0.465
ARIMA (2,1,0)	0.834	0.797	0.691	0.708	0.412
ARIMA (2,1,1)	0.782	0.731	0.703	0.689	0.398
ARIMA (2,1,2)		0.741	0.697	0.695	0.398
ARIMA (2,1,3)		0.772	0.732	0.725	0.385
ARIMA (2,2,0)	1.143	1.119	0.875	0.976	0.537
ARIMA (2,2,1)	0.855				
ARIMA (2,2,2)					

TABLE 8: Best fit of ARIMA orders for each sample (purple cells). At order (1,0,1), the differences (in RMSE) between chosen values (same row as (1,0,1)) and best fit values are relatively small. Grey cells indicate no result.

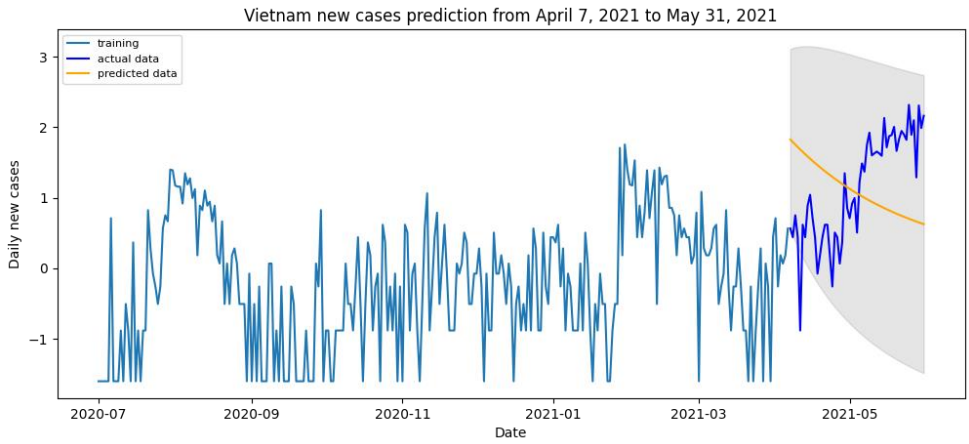


FIGURE 87. Prediction compared to actual data on the last 55 days of dataset.

5.2.7 Observations of the model

According to the methods discussed in Section 5.2.5, six different ARIMA orders that are best suited for different countries were selected. Table x presents final scoring:

Brazil	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.845	0.67	1.002	0.795	0.819	0.8262	1.002	0.67
MAE	0.597	0.566	0.785	0.721	0.7	0.6738	0.785	0.566
R2	-0.296	0.201	-0.097	0.175	0.074	0.0114	0.201	-0.296

US	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.206	0.205	0.288	0.263	0.373	0.267	0.373	0.205
MAE	0.145	0.136	0.197	0.178	0.273	0.1858	0.273	0.136
R2	0.688	0.844	0.602	0.38	0.823	0.6674	0.844	0.38

India	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.091	0.105	1.022	0.115	0.09	0.2846	1.022	0.09
MAE	0.073	0.081	0.394	0.091	0.077	0.1432	0.394	0.073
R2	0.75	0.691	-1.236	0.965	0.955	0.425	0.965	-1.236

Japan	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.318	0.235	0.214	0.318	0.287	0.2744	0.318	0.214
MAE	0.258	0.192	0.172	0.241	0.223	0.2172	0.258	0.172
R2	-0.257	0.844	0.823	0.434	0.373	0.4434	0.844	-0.257

Korea	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.227	0.219	0.272	0.184	0.207	0.2218	0.272	0.184
MAE	0.183	0.175	0.212	0.14	0.169	0.1758	0.212	0.14
R2	0.492	0.944	0.819	0.303	0.143	0.5402	0.944	0.143

Vietnam	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average	Highest	Lowest
RMSE	0.756	0.715	0.7	0.674	0.454	0.6598	0.756	0.454
MAE	0.634	0.591	0.546	0.506	0.374	0.5302	0.634	0.374
R2	-0.148	-0.197	0.332	0.259	0.632	0.1756	0.632	-0.197

TABLE 9. Final scoring with average, highest and lowest values for each country's model.

Average R^2 Score ranging from 0.0114 (Brazil) to 0.6674 (United States). On the other hand, the accumulated error (MAE) from India was the lowest, 0.1432, while for Brazil it was the highest, 0.6738.

Since the R^2 Score is a percentage measure, when analyzing each prediction it is expected to see results closer to 1.0 (100%). This means current results are poor. Higher R^2 values can be obtained if the sample time series increases or decreases gently. The error metrics would be higher if the prediction were performed on oscillating data series. This could suggest that using ARIMA models as a predictive tool for COVID-19 could be appropriate if the country does not face sharp fluctuations in infection numbers. In the case of time series with noticeable changes in the epidemiological curves, the errors tend to increase significantly.

When analyzing the performance of the distance-based error metric MAE Scores, it is important to keep in mind that their results are dependent on the sample values, since the distance is calculated based on the subtraction of the average. That's the reason why the average MAE result for Korea, which achieved the second-worst performance for R^2 Score, would be considered the best average result, since it presents a relatively low value for MAE.

Prediction results for last 55 days in all of the datasets show bad performances. This suggests that further operations need to be done on transforming data and also extended ARIMA models (SARIMA, SARIMAX) should be considered for better forecasting result. It is important to notice that the ARIMA model assumes requirements of stationarity and seasonality that can be approximated by numerical calculations, but are often viewed as weak assumptions. Data should be transformed more properly to achieve true stationarity.

Another key aspect is related to the definition of the model parameters. A detailed analysis of the ACF or PACF behaviors should be carefully conducted to characterize the time series under analysis.

5.3 FBProphet Model

5.3.1 Introduction

FBProphet is an open-source forecasting framework developed by Facebook's data scientist team. [69] It is used for forecasting time series data based on an additive model which can make the task of forecasting more accessible and easier to carry out. It works best with time series that have strong seasonal effects and several seasons of historical data. FBProphet uses ARIMA, exponential models, and other similar regressive models. Historical time series records can be used to forecast the values of the future.

FBprophet is based on decomposable time-series combined in this equation

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t$$

in which:

g(t): represents the trend function; piecewise linear or logical growth to fit non-periodic changes in the value of the time series

s(t): represents the periodic changes as a week and/or year seasonality

h(t): represents the effects of holidays that occur on irregular schedules over a day or more.

εt: represents any unusual change which is not accommodated by the model.

Basic steps of the prediction procedure method of FBProphet are shown below:

1- Daily collect historical time series records

2- Since the dataset is made of daily records, they need to be converted to date instead of a string format. Then the transformation of the Confirmed cases should be applied by logarithmic function to be linear.

3- Based on the historical dataset, the model will be fitted by the framework of FBProphet.

4- According to the forecasting model fitted, a suitable algorithm is generated to predict the defined period.

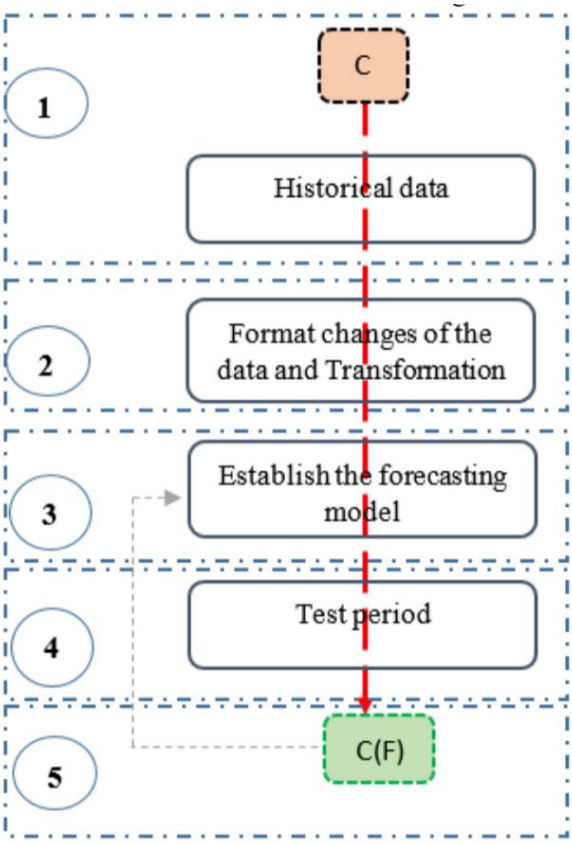


FIGURE 88. FBProphet general prediction procedure

The results depend on the quantity and quality of the dataset injected and fitted into the model to forecast. In particular, we assume that the average frequency and magnitude of trend changes in the future will correspond to those that we observe in the historical time series. [70]

5.3.2 Pre-proposed method

The goal of forecasting with FBProphet is to predict the number of new cases each day for the next 14 days since May 31, 2021, which represents out-of-range forecast. However, we also introduced the dataset of Global active cases when we observed this model in the first place.

Custom records of six remaining regions contain only Date and New cases attributes were added later in the process. Data were transformed using Box-Cox Transformation method as described in Section 5.2.5.2.

The terms R^2 and MAE scores was taken into account in the assessment.

5.3.3 Results

The results are shown in the following order: Global, Brazil, United States, India, Japan, South Korea, Vietnam.

5.3.3.1 Global

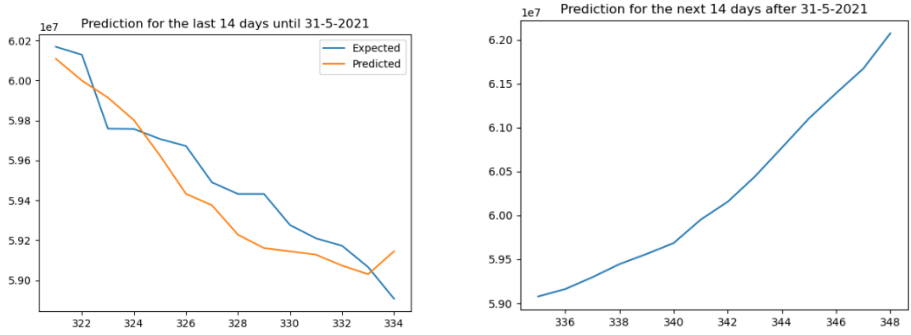


FIGURE 89: Prediction result for last 14 days of the Global Active dataset (left) and prediction of the next 14 days since May 31, 2021 (right).

5.3.3.2 Brazil

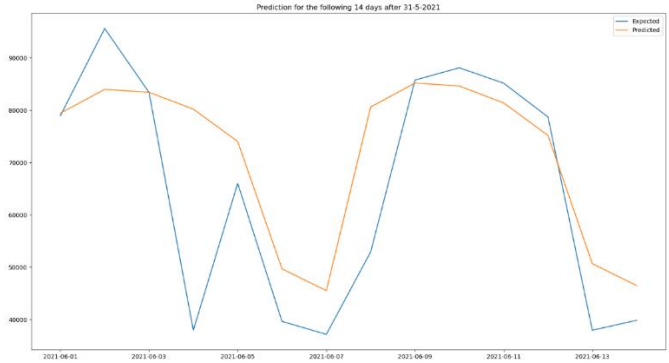


FIGURE 90: Forecasting result for daily new cases of the next 14 days since May 31, 2021, Brazil. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.3.3 United States

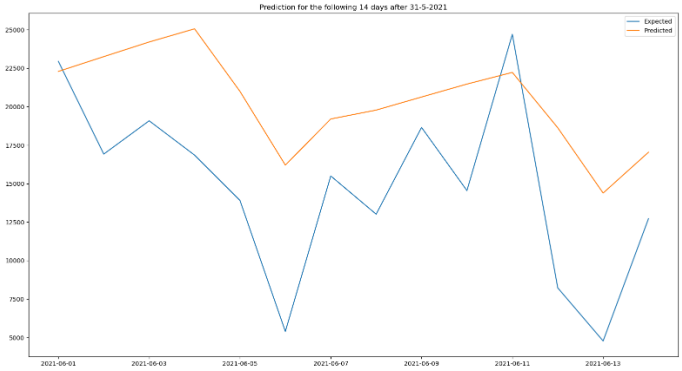


FIGURE 91: Forecasting result for daily new cases of the next 14 days since May 31, 2021, United States. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.3.4 India

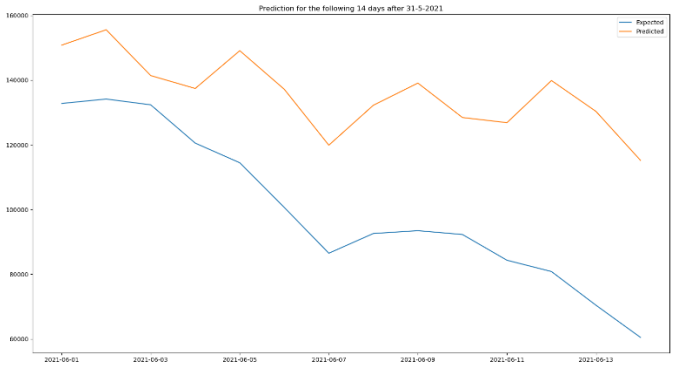


FIGURE 92: Forecasting result for daily new cases of the next 14 days since May 31, 2021, India. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.3.5 Japan

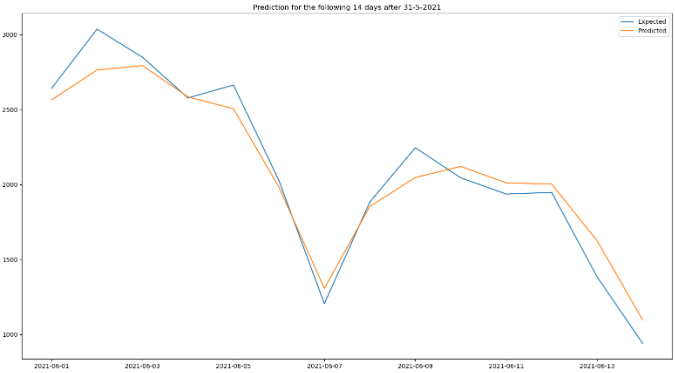


FIGURE 93: Forecasting result for daily new cases of the next 14 days since May 31, 2021, Japan. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.3.2 South Korea

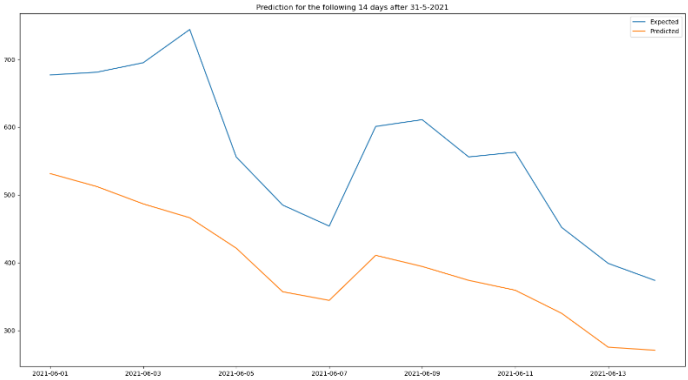


FIGURE 94: Forecasting result for daily new cases of the next 14 days since May 31, 2021, South Korea. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.3.2 Vietnam

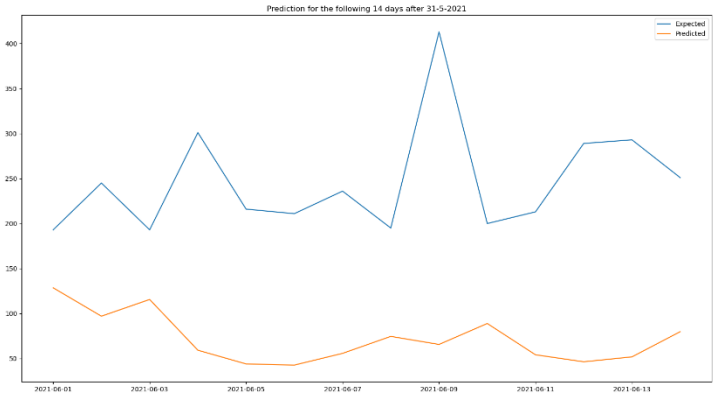


FIGURE 95. Forecasting result for daily new cases of the next 14 days since May 31, 2021, Vietnam. The orange line is the forecasted data line graph while the blue line is the actual data from June 1, 2021 to June 14, 2021.

5.3.4 Observations of the model

Evaluation results are displayed in the table:

	R2	MAE
Brazil	0.024	9914.898
US	-4.078	5945.643
India	-11.654	36292.47
Japan	0.926	110.728
Korea	-3.958	170.983
Vietnam	-51.211	174.584

TABLE 10: A drag-and-drop process in RapidMiner

Most of the R^2 values reached negative, while the MAEs vary depending on the range of data observed. For countries with smaller number of cases such as Vietnam, with the data range from 0 to 400, the MAE score is considered large and therefore forecasting result tends to go wrong. On the other hand, countries with higher records are more likely to have low MAE, which indicates a higher accuracy of the prediction.

FBProphet is certainly a good choice for fast, accurate predictions with intuitive, changeable parameters. While this particular library is not particularly robust, it is fast and gives some very good results for this first pass in predicting time series data.

5.4 RapidMiner Automodel

5.4.1 RapidMiner software

RapidMiner is a free of charge, open-source software tool for data and text mining. It can be used for a wide variety of data and text mining projects.

FIGURE 96. RapidMiner Corporate logo [71]

The standard implementation of procedures like such as data cleansing, visualization, pre-processing can be done with drag-and-drop options without having to write a single line of code.

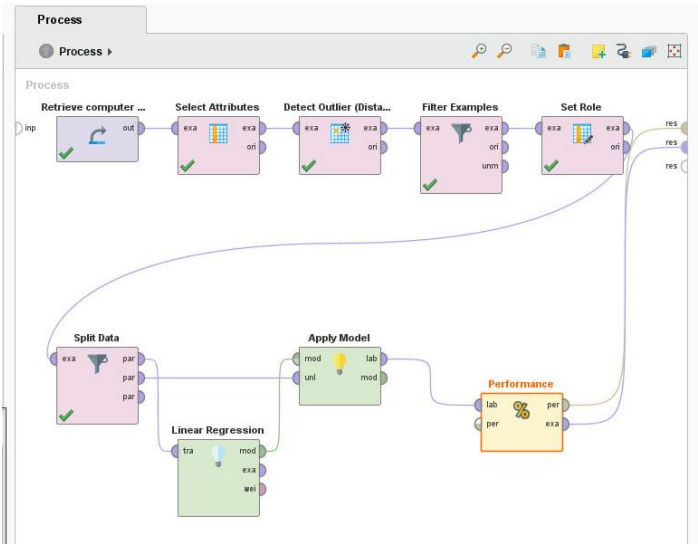


FIGURE 97. A drag-and-drop process in RapidMiner

RapidMiner provides a wide range of machine learning algorithms from classification to clustering and regression. Because of its amazing facilities, users find this tool very useful and easy to use. [72]

5.4.2 RapidMiner Auto Model tool for COVID-19 dataset

Auto Model is an extension to RapidMiner Studio that accelerates the process of building and validating models. [73] In this process, only the dataset of Japan has been used.

Model	Relative Error	Root mean squared error	Absolute error	Relative error lenient	Squared error	Correlation
Gradient Boosted Trees	0.055687537	6187.165 +/- 262.045	4557.255 +/- 181.221	19.87% +/- 0.46%	38335940.390 +/- 3235662.422	0.966 +/- 0.001
Decision Tree	0.081237991	9501.505 +/- 857.883	7080.783 +/- 424.311	27.07% +/- 1.53%	90867374.421 +/- 16236145.882	0.929 +/- 0.012
Random Forest	0.143267063	3431.359 +/- 1010.122	1966.803 +/- 386.474	8.12% +/- 1.12%	12590500.570 +/- 6693870.146	0.989 +/- 0.007
Generalized Linear Model	0.198659166	3819.881 +/- 456.789	2772.169 +/- 212.237	14.33% +/- 0.86%	14758418.371 +/- 3325913.723	0.992 +/- 0.003
Deep Learning	0.270650191	1579.167 +/- 288.263	1095.457 +/- 215.990	5.57% +/- 0.78%	2560245.199 +/- 918347.712	0.998 +/- 0.001
Support Vector Machine	0.386122586	18825.791 +/- 648.423	12610.771 +/- 507.544	38.61% +/- 1.20%	354746768.465 +/- 24676490.093	0.760 +/- 0.011

TABLE 11: Results after running automodel for Japan data sample.

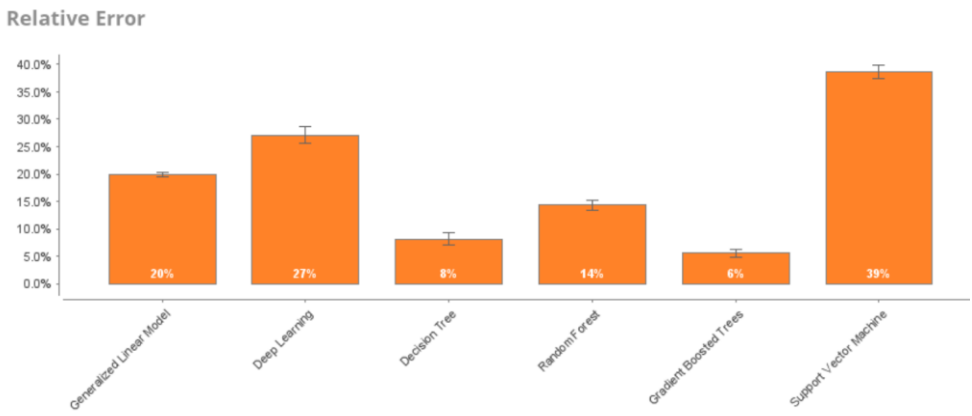


FIGURE 98. Relative error of models

Based on the result table, RFR is a relatively promising choice amongst other models. RFR's errors are minor in all the 6 criteria.

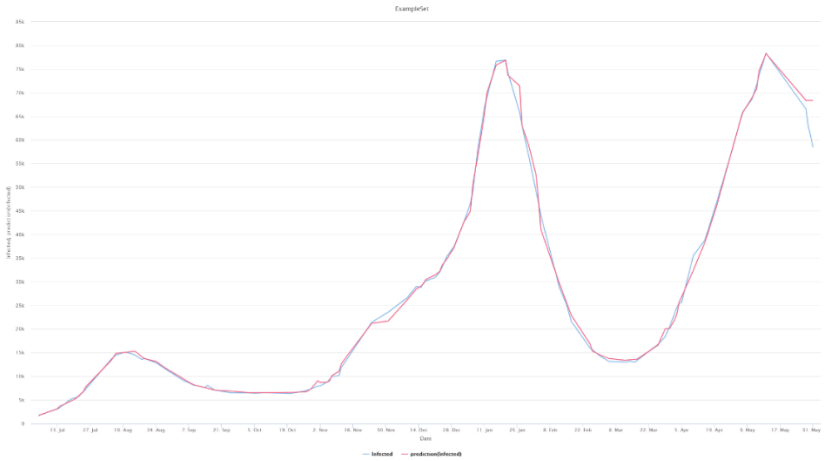


FIGURE 99. Prediction result of separated RFR model in RapidMiner. The blue line is the original data, while the red line is the prediction time series. Performance score R^2 is 1707.057, with data range varies from 1821 to 78346.

We also experienced with ARIMA in RapidMiner. Below is the model's prediction for 5 day forward since May 31, 2021:

Date	Real data	Predicted data
01 June 2021	56257	58060.025152631126
02 June 2021	54436	57989.70890478191
03 June 2021	51915	57919.55329130818
04 June 2021	50019	57849.557945247754
05 June 2021	47887	57779.72250047679

TABLE 12: Real data of Japan with data predicted by ARIMA model

5.5 Other models that we didn't actually take in

5.5.1 Kalman Filtering

The Kalman Filter is a state-space model that is used in several applications as a predictor. The filter algorithm requires low computational power and provides estimates of some unknown variables given the measurements observed over time. However, the mathematical concepts are not so simple to understand.

In practice, the method considers a set of measures observed over an interval, including noise, and estimates new samples, according to the considered time series or variable. The first concept is to understand that it considers a joint probability distribution across the variables for each time frame. To simplify, the Kalman Filter (KF) is an optimization estimator which suggests parameters of interest from previous observations.

The KF aims to find the “most reliable estimate” from noisy input. The filter presents a recursive resolution to the linear optimal filtering problem to stationary as well as nonstationary situations, and treats the new measures as they appear. Only the previous estimate is used for calculation, which reduces the need for saving the whole data from previous iterations. These techniques have found application in various disciplines and, across the past two decades, have been used to contagious infection epidemiology. [74]

5.5.2 Variations of SIR model

5.5.2.1 SIR-D model

The SIR-D model deals with only the number of susceptible, infected, recovered, and dead people. According to this model, a susceptible person in contact with an infected person is prone to get infected. An infected person can either recover from the disease or die due to the infection. Thus, this model considers that the sum of S(t), I(t), R(t), and D(t) remains constant. Further, it is assumed that all people who are exposed to the virus get infected immediately, that is, there is no latent time between the exposure and infection. No effect of confinement or quarantine is considered.

The governing equations of this model are

$$\frac{dS}{dT} = -N^{-1}\beta SI$$

$$\frac{dI}{dT} = N^{-1}\beta SI - (\gamma + \alpha)I$$

$$\frac{dR}{dT} = \gamma I$$

$$\frac{dD}{dT} = \alpha I$$

in which:

S(t)—susceptible population, people who can be infected by the virus at a particular time t.

I(t)—people who have been infected by the virus at a particular time t, that is, active number of cases of infection.

R(t)—cumulative number of people at time t who have recovered from the infection.

D(t)—cumulative number of people at time t who have died due to the infection. [75]

5.5.2.2 SEIR model

The SEIR mathematical model is an extensively used compartmental epidemic model that is based on the division of the population into four basic compartments; an individual can

either be susceptible (S), exposed to the disease but not yet infectious (E), infectious (I), or removed (recovered or deceased) (R). In a closed population without births or deaths, the SEIRS model is:

$$\frac{dS}{dt} = b(N - S) - \frac{\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - (\alpha + b)E$$

$$\frac{dI}{dt} = \alpha E - (\gamma + b)I$$

$$\frac{dR}{dt} = \gamma I - bR$$

in which

S(t) - the fraction of susceptible individuals (those able to contract the disease)

E(t) - the fraction of exposed individuals (those who have been infected but are not yet infectious)

I(t) - the fraction of infective individuals (those capable of transmitting the disease)

R(t) - the fraction of removed individuals (those who have recovered or deceased). [76]

6 Changes compared to the latest version of report

Here are major changes from the latest version (July 10, 2021):

- *The previous research object, Active Cases, relates to the three terms Confirmed, Deaths, and Recovered. Since there is a lack of data on Recovered in many countries, predictions for this criterion can be trivial. So we considered a new target for prediction – Daily New Cases. It is more likely to encounter different time series patterns in such data sets than to observe cumulative attributes.*
- *We introduced a completely new approach to ARIMA model and observe new important terms such as data stationarity or Grid Search CV method for determining the best order. For each dataset, five rolling-forward validation samples are considered, thus improve the overall results. R2 score and MAE score evaluations were also added. The process is now stuck at the back-transformation of predicted data. (results shown in the report are still in transformation form.)*
- *We started experimenting with FBProphet and got amazing results. The FBProphet is a useful tool for forecasting studies. Still, further observations are required to achieve high accuracy predictions.*
- *Added Python-graphed visuals for daily New cases and daily Death reports.*
- *Replace some figures with hand-drawn images for better explaination.*
- *Re-format the report paper.*

7 Conclusions

WHO has declared COVID-19 a pandemic because it has infected most countries and poses a major threat to humanity. In this article we performed an analysis and prediction of the disease using different types of models. We have collected COVID-19 data from 3 severely affected countries Brazil, USA, India; 3 Asian countries Japan, South Korea, Vietnam and worldwide by May 31, 2021 at the latest. In most of the country data, ARIMA has the best performance compared to others on the scale of R^2, MAE and RMSE. The trend analysis shows a rapid growth in confirmed and deceased cases, and the predictive study shows a large increase in expected active and new cases worldwide. However, lockdowns and containment policies can affect the prediction results. The adopted models have proven themselves well, but limit our study to the effectiveness of the models, which can be further improved by an ensemble of several prediction models. The forecast results obtained can be further improved by taking into account various variables such as population density, weather, health system, patient history, etc. using deep learning techniques and artificial intelligence.

The aim of this study is to provide some insights into the COVID-19 pandemic from a data analysis perspective in a didactic and simple way. Predicted results should in no way be taken as confirmation of what will happen in the future. Observations obtained from data exploration are personal opinions.

Acknowledgements

We would like to thank many people who have encouraged and supported us in various ways throughout the process, namely our team, teachers and fellow friends. Special thanks to one of our friends for the great help in getting us familiar with the concept of machine learning in the first few steps.

We are also very grateful to our teacher, Professor Tran Tuan Anh, for this unique and interesting self-exploration researching experience. Looking back, we have certainly come a long way compared to where we started.

Last but not least, our acknowledgements would remain incomplete if we do not thank the authors of related publications and articles on several researchers social platforms (Kaggle, ResearchGate,...). Without your resources, we can never achieve such knowledge and complete the project.

References

[1] [70] [74] [s] Marques, J., Gois,F., Neto,J., Fog,S. (2020). *Predictive Models for Decision Support in the COVID-19 Crisis*. Springer Nature. ISBN 978-3-030-61913-8. https://doi.org/10.1007/978-3-030-61913-8

[2] Wouter G. Touw, Jumamurat R. Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, Sacha A. F. T. van Hijum, *Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?*, Briefings in Bioinformatics, Volume 14, Issue 3, May 2013, Pages 315–326, https://doi.org/10.1093/bib/bbs034

[3] Gebretensae YA, Asmelash D. *Trend Analysis and Forecasting the Spread of COVID-19 Pandemic in Ethiopia Using Box–Jenkins Modeling Procedure*. Int J Gen Med. 2021;14:1485-1498. https://doi.org/10.2147/IJGM.S306250

[4] University of Ottawa. (n.d.). *Useful Concepts: High Risk vs. Population Health Interventions.* http://www.med.uottawa.ca/courses/epi6181/course_outline/Concepts-prev.htm

[5] Entrancei. (n.d.). *Types Of Diseases – Human Health And Diseases of Class 10.* https://www.entrancei.com/chapter-human-health-and-diseases/types-of-diseases

[6] [Figure] Ignasm, M. (2017, March 2). *Malimu intro to epidemiology*. https://www.slideshare.net/MiharbiIgnasm/malimu-intro-to-epidemiology

[7] **MLT Hub with kamran** on Youtube. (2020, October 1). *Important Definitions ||Endemic|| ||Epidemic|| Pandemic || Sporadic disease*. https://www.youtube.com/watch?v=4sQKHAnPPkU

[8] Wikipedia. (n.d.). *Pandemic*. https://en.wikipedia.org/wiki/Pandemic

[9] WebMD. (n.d.). *Pandemics*. https://www.webmd.com/cold-and-flu/what-are-epidemics-pandemics-outbreaks

[10] [Figure] Financial Express. (2021, May 6). *Black Death to Covid-19: A look at the history of pandemics that ravaged the planet.* https://www.financialexpress.com/lifestyle/health/black-death-to-covid-19-a-look-at-the-history-of-pandemics-that-ravaged-the-planet/2246918/

[11] Cartwright, M. (2020, March 28). *Black Death*. World History Encyclopedia. https://www.worldhistory.org/Black_Death/

[12] Flatley, L. (2020, April 30). *COVID-19 Pandemic is a Teachable Moment on Native American History.* https://www.eriereader.com/article/covid19-pandemic-is-a-teachable-moment-on-native-american-history

[13] Claeson, M. (n.d.). *Cholera*. https://www.britannica.com/science/cholera

[14] Taubenberger JK, Morens DM. *1918 Influenza: the mother of all pandemics*. Emerg Infect Dis. 2006;12(1):15-22. DOI:10.3201/eid1201.050979

[15] Centers for Disease Control and Prevention (CDC). (n.d.). *1957-1958 Pandemic (H2N2 virus)*. https://www.cdc.gov/flu/pandemic-resources/1957-1958-pandemic.html#:~:text=The%20estimated%20number%20of%20deaths,116%2C000%20in%20the%20United%20States.

[16] Sino Biological. (n.d.). *Hong Kong Flu (1968 Influenza Pandemic)*. https://www.sinobiological.com/research/virus/1968-influenza-pandemic-hong-kong-flu

[17] Centers for Disease Control and Prevention (CDC). (n.d.). *HIV Basics – Basic Statistics.* https://www.cdc.gov/hiv/basics/statistics.html#:~:text=About%2037.9%20million%20people%20were,the%20start%20of%20the%20epidemic.

[18] Bishop, J. (2020, June 18). *Economic Effects of the Spanish Flu*. https://www.rba.gov.au/publications/bulletin/2020/jun/economic-effects-of-the-spanish-flu.html

[19] Ceylan, R. F., Ozkan, B., & Mulazimogullari, E. (2020). *Historical evidence for economic effects of COVID-19*. The European journal of health economics : HEPAC : health economics in prevention and care, 21(6), 817–823. https://doi.org/10.1007/s10198-020-01206-8

[20] Zhu, Z., Lian, X., Su, X. et al. *From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses*. Respir Res 21, 224 (2020). https://doi.org/10.1186/s12931-020-01479-w

[21] Wang, M. D., & Jolly, A. M. (2004). *Changing virulence of the SARS virus: the*

epidemiological evidence. Bulletin of the World Health Organization, 82(7), 547–548.

[22] Arabi, Y.M., Balkhy, H.H., Hayden, F.G., Bouchama, A., Luke, T., et al. (2017). *Middle East Respiratory Syndrome*. N Engl J Med 2017; 376:584-594. DOI: 10.1056/NEJMsr1408795

[23] Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., Montgomery, J. M., … Widdowson, M. A. (2012). *Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study*. The Lancet. Infectious diseases, 12(9), 687–695. https://doi.org/10.1016/S1473-3099(12)70121-4

[24] World Health Organization (WHO). (2016, April 15). *Ebola Situation Reports*. https://apps.who.int/ebola/ebola-situation-reports

[25] Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. Acta bio-medica : Atenei Parmensis, 91(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

[26] [Figure] Fabex. (2020, June 2). Coronavirus COVID-19. https://fabex.fr/en/coronavirus-covid-19/

[27] World Health Organization (WHO). (2020, February 16-24). *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) (pdf).* https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)

[28] Hamzah, B., Fairoza Amira & Hau, Cher & Nazri, Hafeez & Ligot, Dominic & Lee, Guanhua & Shaib, Mohammad & Zaidon, Ummi & Abdullah, Adina & Chung, Ming & Ong, Chin & Chew, Pei. (2020). *CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction*. 10.2471/BLT.20.255695. https://www.researchgate.net/publication/340032869_CoronaTracker_World-wide_COVID-19_Outbreak_Data_Analysis_and_Prediction

[29] Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). *COVID-19 Data Repository*. Github. https://github.com/CSSEGISandData/COVID-19

[30] Dong E, Du H, Gardner L. (2020, February 19). *An interactive web-based dashboard to track COVID-19 in real time*. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1"

[31] Our World in Data (OWID). (n.d.). *COVID-19 Dataset*. Github. https://github.com/owid/covid-19-data

[32] Google. (n.d.). *COVID-19 Public Datasets*. https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-public-data-program

[33] Shinde, G.R., Kalamkar, A.B., Mahalle, P.N. et al. (2020, June 11). *Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art*. SN COMPUT. SCI. 1, 197 (2020). https://doi.org/10.1007/s42979-020-00209-9

[34] Python Software Foundation. *Python Official Website*. https://www.python.org/

[35] NumFOCUS. *pandas Official Website*. https://pandas.pydata.org/

[36] The Matplotlib development team. *Matplotlib: Visualization with Python*. https://matplotlib.org/

[37] Waskom,M. *seaborn: statistical data visualization*. https://seaborn.pydata.org/

[38] Microsoft. *Microsoft Excel*. https://www.microsoft.com/en-ww/microsoft-365/excel

[39] IBM. *IBM SPSS Statistics*. https://www.ibm.com/products/spss-statistics

[40] Robbins, N. (2012, January 4). A Histogram is NOT a Bar Chart. Forbes. https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=d6baea86d775

[41] [Figure] Freeman, J. (2021, April 3). *Bar Chart vs Histogram*. https://www.edrawsoft.com/histogram-vs-bar-chart.html

[42] ASQ. (n.d.) WHAT IS A HISTOGRAM? https://asq.org/quality-resources/histogram

[43] PQ System. (n.d.) Histogram: Study the shape. https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/interpretation/histogram_shape.php

[44] Worldometer. (n.d.) *India Population*. https://www.worldometers.info/world-population/india-population/

[45] Sy, K.T.L, White, L.F, Nichols, B.E. (2021). *Population density and basic reproductive number of COVID-19 across United States counties*. PLOS ONE 16(4): e0249271. https://doi.org/10.1371/journal.pone.0249271

[46] DEKA (dkjung on Kaggle). (n.d.). *[COVID-19] EDA (S.Korea) | Forecasting (Global)*. https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global/data

[47] Our World in Data. *COVID-19: Stringency Index*. https://ourworldindata.org/grapher/covid-stringency-index

[48] FT Visual & Data Journalism Team. (2021, June 24). *Lockdowns compared: tracking governments' coronavirus responses*. Financial Times. https://ig.ft.com/coronavirus-lockdowns/

[49] Wikipedia. (n.d.). *Susceptible individual*. https://en.wikipedia.org/wiki/Susceptible_individual

[50] [Figure] Encyclopaedia Britannica. (n.d.). [*Figure of Exponential growth and Logistic growth], Carrying capacity*. https://www.britannica.com/science/carrying-capacity

[51] **evenabrokenclock10** on Even a Broken Clock. (2017, March 28). *Exponential Decay Curve in Politics.* https://evenabrokenclock.blog/2017/03/28/exponential-decay-curve-in-politics/

[52] Salathé, M & Case, N. (2020, May). *What Happens Next?* https://ncase.me/covid-19/

[53] (n.d.) *The SIR epidemic model*. https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/

[54] Bakshi, C. (2020, June 9). *Random Forest Regression.* Medium. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

[55] scikit-learn. *sklearn.ensemble.RandomForestRegressor*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[56] Sharma, A. (2020, May 12). *Decision Tree vs. Random Forest – Which Algorithm Should you Use?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

[57] Australian Bureau of Statistics. (n.d). *Time Series Analysis: The Basics.* https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics

[58] [61] Hyndman, R.J., & Athanasopoulos, G. (2018). Chapter 2.3: Time series patterns, *Forecasting: principles and practice, 2nd edition*, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on July 17, 2021.

[59] influxdata. (n.d.) *What is time series data?* https://www.influxdata.com/what-is-time-series-data/

[60] NIST/SEMATECH (2003). Chapter 6 Process or Product Monitoring and Control, 6.4.4.2. Stationarity,.*e-Handbook of Statistical Methods*. https://doi.org/10.18434/M32189

[62] Prabhakaran, S. (2019, February 18). *ARIMA Model – Complete Guide to Time Series Forecasting in Python*. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

[63] Corporate Finance Institute (CFI). (n.d.). *What is Autocorrelation?* https://corporatefinanceinstitute.com/resources/knowledge/other/autocorrelation/

[64] Singh, A. (2018, September 13). *A Gentle Introduction to Handling a Non-Stationary Time Series in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/non-stationary-time-series-python/

[65] Glen, S. (2015, July 14). *What is a Box Cox Transformation*? From StatisticsHowTo.com. https://www.statisticshowto.com/box-cox-transformation/#:~:text=A%20Box%20Cox%20transformation%20is,a%20broader%20number%20of%20tests.

[66] Plummer, A. (2020, September 2). *Box-Cox Transformation: Explained*. Towards data science, Medium. Accessed on July 17, 2021. https://towardsdatascience.com/box-cox-transformation-explained-51d745e34203

[67] Scipy library in Python. *scipy.stats.boxcox documentation*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html

[68] Brownlee, J. (2016, December 19). *How To Backtest Machine Learning Models for Time Series Forecasting*. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/

[69] Facebook. *FBProphet Official Website*. https://facebook.github.io/prophet/

[71] RapidMiner.Corp. *RapidMiner Official Website*. https://rapidminer.com/

[72] Kotu, V., Deshpande, B. (2019). *Chapter 15 - Getting Started with RapidMiner, Data Science (Second Edition).* ISBN 9780128147610. 491-521. https://doi.org/10.1016/B978-0-12-814761-0.00015-0. https://www.sciencedirect.com/science/article/pii/B9780128147610000150

[73] RapidMiner. Auto Model Documentation. (n.d) https://docs.rapidminer.com/9.5/studio/guided/auto-model/

[75] Sen, D., & Sen, D. (2021). Use of a Modified SIRD Model to Analyze COVID-19 Data. Industrial & Engineering Chemistry Research, acs.iecr.0c04754. https://doi.org/10.1021/acs.iecr.0c04754 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7875346/

[76] Ajbar, A., Alqahtani, R.T. Bifurcation analysis of a SEIR epidemic system with governmental action and individual reaction. Adv Differ Equ 2020, 541 (2020). https://doi.org/10.1186/s13662-020-02997-z