# Statistical Modeling of Daily Confirmed COVID-19 Cases and Deaths in Europe and United States

Zerui Zhang
*Marquette University*

# STATISTICAL MODELING OF DAILY CONFIRMED COVID-19 CASES AND DEATHS IN EUROPE AND UNITED STATES

by

Zerui Zhang

A Thesis Submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

August 2021

# ABSTRACT

## STATISTICAL MODELING OF DAILY CONFIRMED COVID-19 CASES AND DEATHS IN EUROPE AND UNITED STATES

Zerui Zhang

Marquette University, 2021

A novel coronavirus disease was first discovered in Wuhan, China, in December 2019. This new coronavirus named COVID-19 has rapidly spread and become a global threat affecting almost all the countries in the world. Therefore, it is important to know the trend of coronavirus disease to mitigate its effects. A good prediction model is crucial for the health care system to understand the trend of the COVID-19.

This study aims to construct a good prediction model. Firstly, we detect change points of the time series data of daily confirmed cases and deaths of COVID-19 in the United States and Europe, and secondly, construct prediction models for daily confirmed cases and deaths of COVID-19 based on the data that was divided by the change points, and thirdly, select the best prediction model to forecast the future number of daily confirmed cases and deaths of COVID-19 in the United States and Europe.

The data was collected from the official website of the Centers for Disease Control (CDC) and Our World in Data from August 1st, 2020 to January 23th, 2021, and we used daily confirmed cases and deaths of COVID-19 in the United States and Europe. An Auto-Regressive Integrated

Moving Average (ARIMA) model was used to predict the daily new confirmed cases and deaths of COVID-19 from January 24th, 2021 to February 22th, 2021.

This study finds that Change-Point ARIMA models that was divided the data by change points improve the forecasting trends of daily new confirmed cases and deaths of COVID-19 in the United States and Europe.

# ACKNOWLEDGEMENTS

Zerui Zhang

I would like to thank my thesis advisor, Dr. Bansal for his advice and help. I would like to also thank committee members for careful reading of the thesis and for suggesting helpful changes to improve the thesis. I would also like to thank my parents they support me to go to study abroad and give me a chance to be better. I am also grateful for the support of my whole family.

# Table of Contents

# Chapter 1: Introduction

A new coronavirus disease was first discovered in Wuhan, China, in December 2019. This new virus belongs to the Corona virus's family and thus it was named 2019-nCoV, popularly known as COVID-19. This virus can lead humans to illness and death. The common symptoms of this new coronavirus disease include fever, cough, fatigue, breathing difficulties, and loss of smell and taste. The COVID-19 was identified as a zoonotic coronavirus that means the virus was transferred to humans from animals. It is similar to Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) that is transferred to humans from humans and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) that is transferred to humans from infected dromedary camels. The coronavirus disease has rapidly spread and has infected more than 100 million people, with almost 3,000,000 deaths as of March 2021 around the world. Although vaccines can now be used to prevent the COVID-19, there are still tens of thousands of new confirmed cases every day in the world. In this case, preventing and preparing healthcare services is very crucial. Modeling and future forecast of the daily number of new confirmed cases and deaths are important to help the treatment system in providing healthcare services for newly confirmed patients. Therefore, the statistical prediction models could be beneficial and meaningful in forecasting and controlling this global pandemic disease threat.

Modeling and forecasting the daily confirmed cases and deaths due to COVID-19 can help and provide healthcare services more information about the number of daily new confirmed cases and deaths so that the services staff can pre-prepare relative armamentarium and safeguard procedures. Therefore, building useful ARIMA models to predict the daily confirmed cases and deaths of COVID-19 is very crucial. However, due to the high volatility of data, simply using these data to fit ARIMA models cannot obtain a good model.

In this study, Auto-Regressive Integrated Moving Average (ARIMA) model was used to predict the daily new confirmed cases and deaths of COVID-19 from January 24th, 2020 to February 22th, 2021. A new method is introduced by setting the change points in the data that split the data into segments. Then selecting the best ARIMA model based on the Akaike Information Criterion (AIC) is introduced which means the model with the smallest value of AIC is selected.

Setting change points played a significant role in fitting Auto-Regressive Integrated Moving Average (ARIMA) models. For the United States and Europe, the best ARIMA model is identified, and then 30 future days are forecasted. The best ARIMA model selection is based on Akaike Information Criterion (AIC), Box-Cox test (compared the p-value), and Residual diagnostics (normality and stationary in variance) is used to validate the Change-Point ARIMA models. Comparing the ARIMA models fitted with change point and the ARIMA models fitted without change point showed that the change point significantly improved the ARIMA model to forecast the daily new confirmed cases and deaths due to COVID-19. The predictive models can provide treatment system for future daily new confirmed cases and deaths of COVID-19 and help healthcare services to prepare newly confirmed patients. The daily confirmed cases and deaths of COVID-19 were collected from the official website of Centers for Disease Control (CDC) and Our World In Data from August 1st, 2020 to January 23th, 2021 and were used to build these ARIMA models.

The rest of this thesis is divided as follows. The second chapter is the theoretical background which includes the theoretical knowledge of the Auto-Regressive Integrated Moving Average (ARIMA) model and Akaike Information Criterion (AIC). The third chapter shows the methodology for building ARIMA models, statistical analysis, results, and discussion in detail. The last chapter gives a conclusion for this thesis.

# Chapter 2: Theoretical Background

In this chapter, the theoretical background of Auto-Regressive Integrated Moving Average (ARIMA) model will be reviewed. This is only a review of ARIMA. Content of this can be found in most of the Time Series books.

For ARIMA models, there are two types of ARIMA models; one is a non-seasonal ARIMA model, denoted as ARIMA (p, d, q). In general, a non-seasonal ARIMA model can be written as $y_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} + e_t$, where $y_t$ is the difference sequence of consecutive observations, p is the order of the autoregressive component, d is the order time series forming a stationary difference process, and q is the order of the moving average component. The other one is a seasonal ARIMA model, denoted as ARIMA (p, d, q) x (P, D, Q)$_m$. Where the first part is the non-seasonal part of the model, the second part is the seasonal part of the model, where P represents the seasonal autoregressive lag, D represents the degree of seasonal differences, Q represents the seasonal moving average and m is the number of cycles.

In general, before we fit the ARIMA model, it is necessary to check whether any time series process is stationary. By the definition, a stationary time series is a series whose statistical characteristics do not depend on the time of the observation series [1]. In a weak sense, a stationary time series means that it has a constant mean and a constant variance. How to make non-stationary time series stationary is the key to analyzing time series. Generally, it is achieved by taking differences. The first difference can be written as $y_t' = y_t - y_{t-1}$. The second-order differencing can be written as $y_t'' = y_t' - y_{t-1}' = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$ and so on.

Differences is also made in the seasonal ARIMA models. The seasonal differencing can be written as $y_t' = y_t - y_{t-m}$. Where m is the number of cycles. The second-order seasonal differencing can be written as $y_t'' = y_t' - y_{t-m}' = (y_t - y_{t-m}) - (y_{t-1} - y_{t-m-1}) = y_t - y_{t-1} - y_{t-m} + y_{t-m-1}$. More clearly, the first difference is the change between one observation and the next observation, and the seasonal difference is the change between an observation and cyclically the next observation.

There is an important notation we need to know when we deal with time series lag, the backshift operator B. It is a useful to symbolic defined as: $By_t = y_{t-1}$, i.e. B operating on $y_t$, which has the effect of shifting the data back by one cycle. When using the backshift operator, the first-order difference can be written as $y_t' = y_t - y_{t-1} = y_t - By_t = (1-B)y_t$. Similarly, the second–order difference can be written as $y_t'' = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1-B)^2 y_t$. In general, the x-order difference can be written as $(1-B)^x y_t$. For the seasonal difference, it can be written as$(1-B)(1-B^m)y_t = (1 - B - B^m + B^{m+1})y_t = y_t - y_{t-1} - y_{t-m} + y_{t-m-1}$.

Differencing order is only one component of the ARIMA models, the other two components of ARIMA models will be discussed in the following paragraph. One is autoregressive component. In an autoregressive model, we use a linear combination of past values of variables to make a prediction. The term autoregression indicates that it is a lagged regression of a variable to itself. Therefore, a p-order autoregressive model can be written as $y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$. Where $\{\varepsilon_t\}$ is white noise. We call it the AR(p) model, which is an autoregressive model of order p.

The other one is moving average component. The moving average model does not use the past value of the predictor variable in the regression but uses the past prediction errors in the quasi-regression model. It can be written as: $y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$. Where $\{e_t\}$ is white noise. We call this equation the MA(q) model, a q-order moving average model. Since we have not observed the value of $e_t$, it is not a regression in the usual sense.

How to select the order of the ARIMA models is also an important part to build an ARIMA model. For the non-seasonal ARIMA model, we select the initial parameters of the non-seasonal ARIMA model from the autocorrelation function (ACF) graph and the partial autocorrelation (PACF) graph. The choice of order p is the lag value after PACF is 0 for the first time, and the choice of q is the lag time after ACF is 0. Empirically, the lower limit and upper limit ACF graph and PACF graph and graph are obtained by testing the null hypothesis related to zero correlation or 0 partial autocorrelation. Therefore, p is selected as the first lag, after which PACF stays within these limits, q is selected as the first lag, after which ACF stays within these limits.

For seasonal ARIMA models, we also select parameters from autocorrelation function (ACF) graphs and partial autocorrelation (PACF) graphs. The order of P is the seasonal lag of the PACF chart, which peaks at the seasonal lag. The selection of the Q order is the same as the selection of the P order, which is to select the seasonal lag peak from the ACF chart.

We also use the Akaike Information Criteria (AIC) to determine the order of an ARIMA model and Akaike Information Criteria (AIC) has been widely used to determine the performance of models. It can be written as $AIC = -2log(L) + 2(p + q + k + 1)$, where L is the likelihood of the data, $p$ is the order of the auto-regressive part and $q$ is the order of the moving average part. The k represents the intercept of the ARIMA models. For AIC, if k = 1, there is an intercept in the ARIMA model (c $\neq$ 0), and if k = 0, there is no intercept in the ARIMA model (c = 0). For the ARIMA models, the corrected AIC can be written as $AIC_c = AIC + \frac{2(p+q+k+q)(p+q+k+2)}{T-p-q-k-2}$.

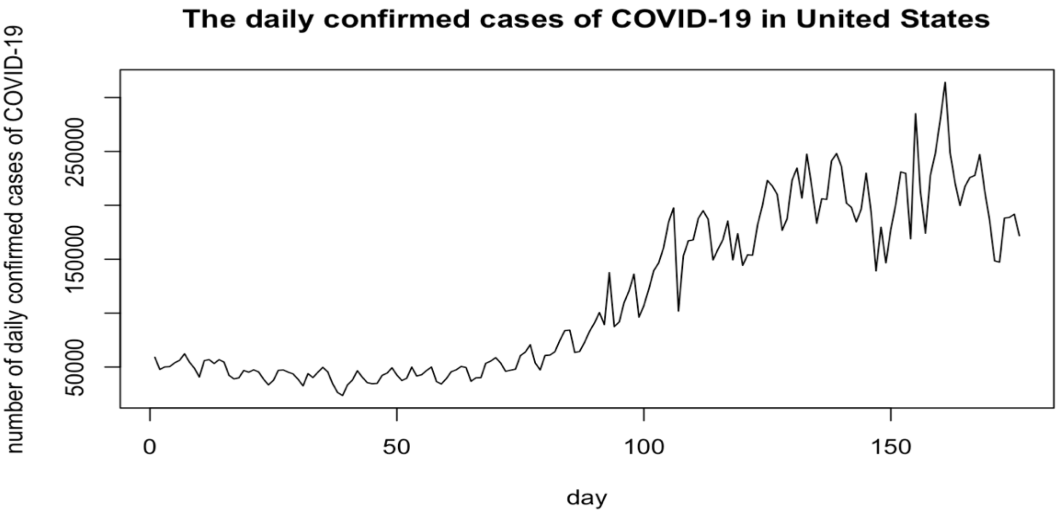
After fitting an appropriate ARIMA model, we can now use ARIMA model to make a prediction. We now discuss how ARIMA model is used to forecast future values. The following three steps can be used to forecast the future values. First, we expand the ARIMA equation so that $y_t$ is on the left side and all other terms are on the right. Second, rewrite the ARIMA equation by displacing t with T + h (a future time point). Finally, on the right side of the ARIMA equation, replace future observations with its past predicted values, replace future errors with zeros, and replace past errors with the corresponding residuals. We will do these three steps in a loop of h starting from 1 to any number of the future time point.

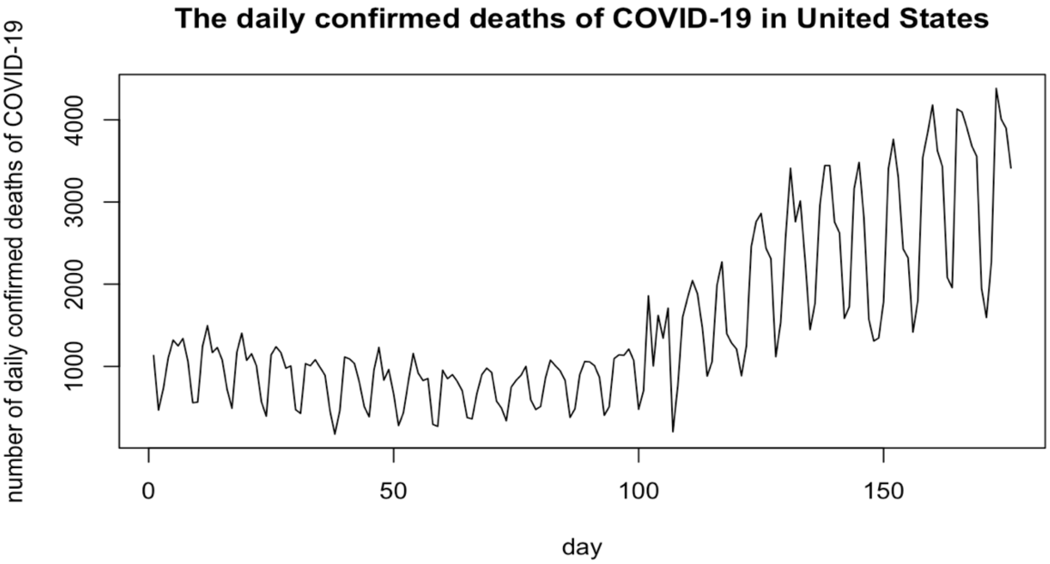
# Chapter 3: Methodology and Results

As we discussed in the previous chapters, modeling future forecast of the daily number of new confirmed cases and deaths is important to help the treatment system in providing healthcare services for newly confirmed patients. Therefore, building useful statistical prediction models and applying these prediction models to forecast the daily number of new confirmed cases and deaths is very important. In this chapter, we will discuss the dataset that was used to build ARIMA models, and the results of the models using R software package. Additionally, the model selection and comparison, the analysis of the model, and the forecasting of the model will be also considered.

## DATASET

The dataset we used in this study came from the official website of the Centers for Disease Control (CDC) [2] and Our World In Data [3] from August 1st, 2020 to January 23th, 2021. The dataset includes the daily confirmed cases and deaths of COVID-19 in the United States and Europe. There is a total of four columns and each column contains 176 days. Figure 1 shows below consisted of four graphs, which plot the daily confirmed cases and deaths of COVID-19 in the United States and Europe from August 1st, 2020 to January 23th, 2021.


The daily confirmed cases of COVID-19 in United States
number of daily confirmed cases of COVID-19
50000
150000
250000
0
50
100
150
day


The daily confirmed deaths of COVID-19 in United States
number of daily confirmed deaths of COVID-19
1000
2000
3000
4000
0
50
100
150
day

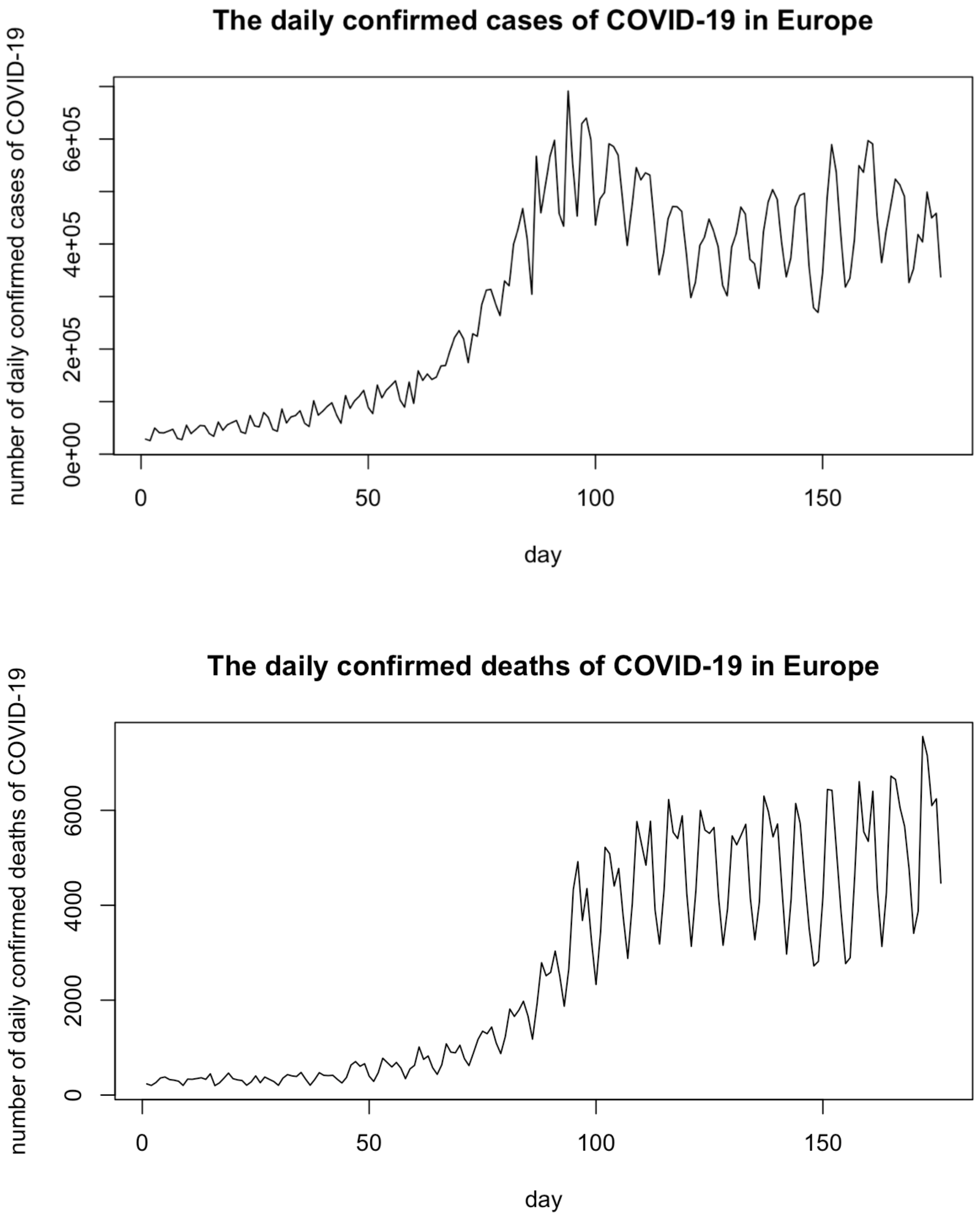


**Figure 1**. The graph of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe from August 1st, 2020 to January 23th, 2021.

As it can be seen from the four graphs in Figure 1, there is a fluctuation trend in both number of cases and number of deaths, but the main trend is increasing. There is one thing that exists in both four graphs that both of the four graphs contain a "jump" point that split the curve into two parts in the graphs. This "jump" point can be identified by observing the fluctuation trend and it is the change points. Finding a change point is important for building the prediction models since it could help to improve the accuracy of the prediction models. It can also help in exploring what caused the change. For example, scientists can explore whether the change occurred due to different variants of COVID-19. Later, will be consider a comparison between the model fitted by the data with change points and the model fitted by the data without change points.

However, a good prediction model cannot be obtained if we use the change point that is only selected by observing graphs to construct a prediction model. Therefore, in this study, we should select a couple of the change points to divide the dataset and use these data to construct the prediction models. Finally, determine which of these predictive models is the best model and use that predictive model to make predictions.

### Change points selection

In this study, the initial change point can be simply selected by observing the "jump" point in the graphs, and then extend 10 points before and back the initial point as change points set (each point means one day). Hence, we consider a set of 21 possible change points that can be used to divide the data. By observing the four graphs in figure 1, the initial change point of these four data can be easy to determine. The initial change point of the daily confirmed cases of COVID-19 in the United States is 92, the initial change point of the daily confirmed deaths of COVID-19 in the United States is 100, the initial change point of the daily confirmed cases of COVID-19 in Europe is 62 and the initial change point of the daily confirmed deaths of COVID-19 in Europe is 93. Table 1 describes all the change points of the data of the daily confirmed cases and deaths of COVID-19 in the United States and Europe.

| Data | Change points |
|---|---|
| Daily confirmed cases in United States | 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102 |
| Daily confirmed deaths in United States | 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110 |
| Daily confirmed cases in Europe | 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72 |
| Daily confirmed deaths in Europe | 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103 |

**Table 1**. All the change points of the data of the daily confirmed cases and deaths of COVID-19 in United States and Europe.

## STATISTICAL ANALYSIS

All analyses were performed using R statistical software, with several different packages used for the main analysis. For constructing ARIMA models, function arima() in the "TSA" package (Kung-Sik Chan, Brian Ripley 2018) was used, which is available from the R Studio version 1.2. The function ACF() and PACF() were used to produces the plot of the ACF and PACF on the same scale and help us to select the parameters of the ARIMA models. These two functions were found in the "stats" package. We also used the function Box.test() in "stats" package, which is used to compute the Box-Pierce test statistic for examining the null hypothesis if the residuals are white noise, which indicates that the ARIMA model is fitted well. . To make Q-Q plot of the residuals to check the goodness of the model, we used the qqnorm() in the package "stats". In the package "stats", we also used the predict() to forecast the future value of the fitted models, and the AIC() was used to compute the value of AIC of the fitted models. We used the function ts.plot() to make a forecasting plot of the fitted ARIMA models. A For Loop was also used to compare the value of the AIC of the 21 fitted models.

**Model selection and comparison**

We used Auto-Regressive Integrated Moving Average (ARIMA) model to forecast the future 30 day of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe. As we introduced in previous chapters, we used the change points set to split the data into two parts, and then used the second part ahead of the change point to fit the ARIMA models, we called the second part of data as new data. We used a total of 21 candidate change points and built 21 ARIMA models. How to know which ARIMA model is the best that is an important step? Hence, in order to select the best ARIMA model, we used Akaike Information Criterion (AIC). The Akaike Information Criterion (AIC) has been widely used in literature to determine the performance of the ARIMA models, and it was the single number score that can be used to judge which is the optimal ARIMA model for a given dataset. It estimates the relative quality of the ARIMA models and only can be used in the comparison with other AIC values of the other ARIMA model. The better model is the one with a lower AIC value. We compared the AIC value of each ARIMA model that was built by fitting the new data. Figure 2 shows the AIC value of each ARIMA model that was fitted by new data.

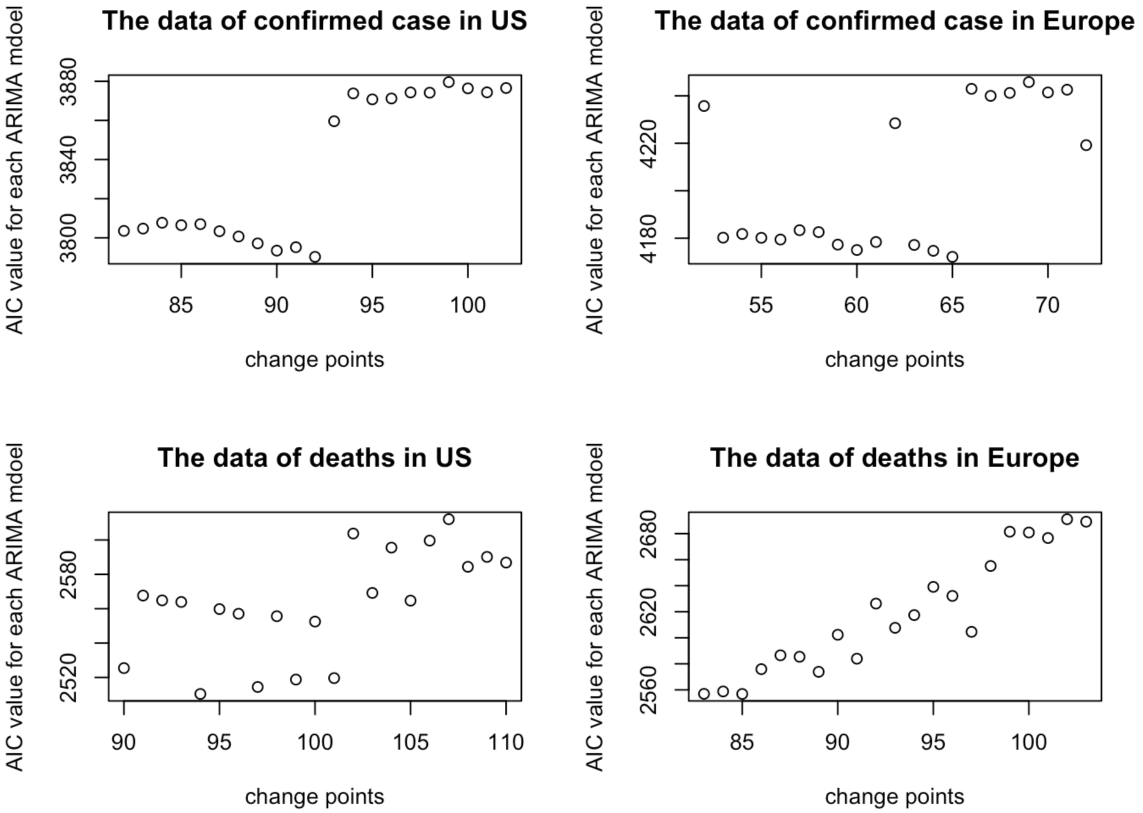


**Figure 2**. The AIC value of each ARIMA model was fitted by new data.

Figure 2 shows all AIC values of each ARIMA model fitted by the new data. Hence, we can obtain the best ARIMA model by seeing the lowest AIC value. In the first plot, the lowest AIC value is at the No. 92 change point that represents October 31, 2020 for the daily new confirmed cases in the United States. In the second plot, the lowest AIC value is at the No. 65 change point that represents October 4, 2020 for the daily new confirmed cases in the Europe. In the third plot, the lowest AIC value is at the No. 94 change point that represents November 2, 2020 for the daily new confirmed deaths in the United States. In the last plot, the lowest AIC value is at the No. 85 change point that represents October 24, 2020 for the daily new confirmed deaths in the Europe. As a result, we used these the best change point to split the data and built ARIMA models by fitting these new data. Hence, the best ARIMA model was selected with these steps.

For selecting the best ARIMA model, the important step is using the change points to obtain the new data, and then use these new data to fit the ARIMA model. We also use the residual diagnostics to prove that using change points to make new data can help to build a better ARIMA model. We also do a comparison of the residuals plot of the ARIMA model fitted by new data and the ARIMA model fitted by original data in Figure 3 and Figure 4.

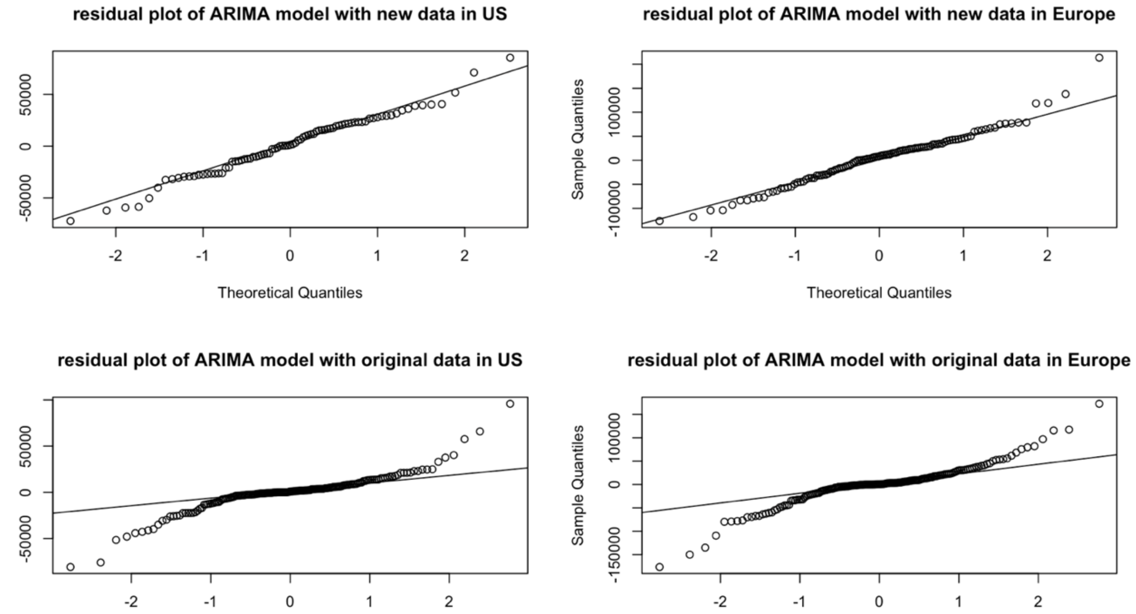


**Figure 3**. The residuals plot of the ARIMA model fitted by new data and of the ARIMA model fitted by original data.

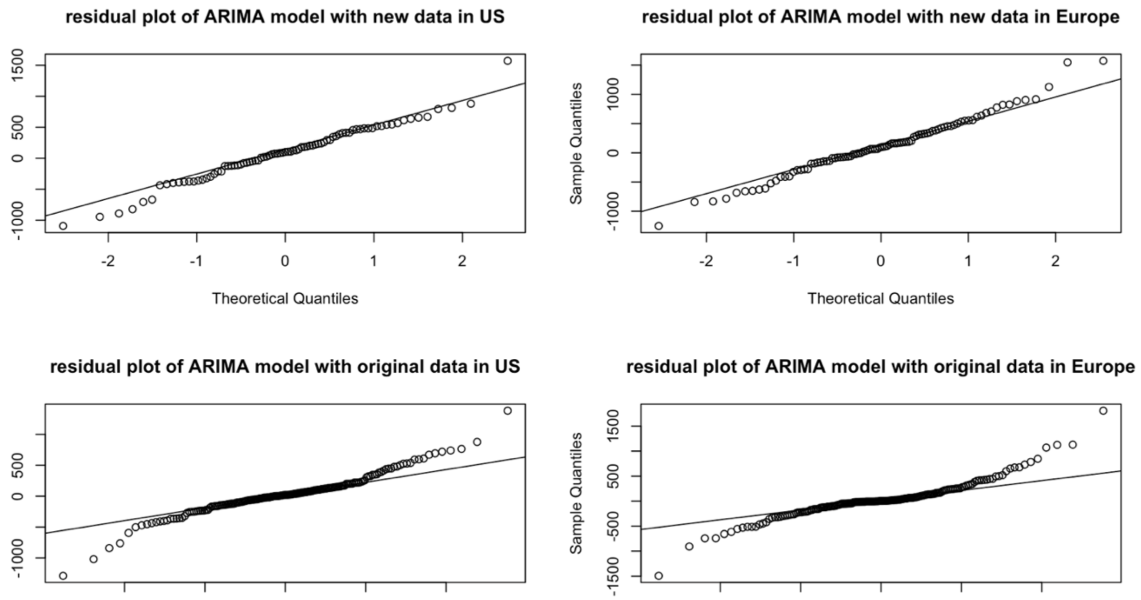


**Figure 4**. The residuals plot of the ARIMA model fitted by new data and of the ARIMA model fitted by original data.

Figure 3 shows the normal probability plots of the residuals of the ARIMA models that was used to forecast the number of daily confirmed cases of COVID-19 in the United States and Europe, and Figure 4 shows the normal probability plots of the residuals of the ARIMA model was used to forecast the number of daily deaths of COVID-19 in the United States and Europe.

By seeing Figure 3 and Figure 4, we can see that the residuals of the ARIMA model fitted by new data look more normal than the residuals of the ARIMA model fitted by original data. All of quantile residuals are almost in a straight line for the new data, and there is an obvious curve in the residual plots of the ARIMA models that were fitted by the data without change points. In this way, we conclude that selecting the best ARIMA model the important step is using the change points to get the new data, and then use these new data to fit the ARIMA model. Therefore, we obtained the best ARIMA model for forecasting the future 30 days of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe.

## RESULTS

In the model selection part, we obtained the best ARIMA model to forecast the future one month of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe. Table 2 shows these four ARIMA models with AIC values.

| ARIMA model used to forecast | ARIMA model | AIC |
|---|---|---|
| Daily confirmed cases in US | $(0, 1, 1)x(0, 0, 1)_7$ | 1969.44 |
| Daily confirmed cases in Europe | $(2, 1, 2)x(0, 0, 1)_7$ | 2748.47 |
| Daily confirmed deaths in US | $(1, 0, 0)x(1, 0, 1)_7$ | 1266.85 |
| Daily confirmed deaths in Europe | $(1, 0, 1)x(1, 0, 1)_7$ | 1429.13 |

Table 2. The best ARIMA model to forecast the future one month of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe, and AIC values of models.

The ARIMA model of forecasting the future 30 days of the number of daily confirmed cases of COVID-19 in the United States is ARIMA $(0, 1, 1)x(0, 0, 1)_7$, the AIC score of it is 1969.44. The ARIMA model of forecasting the future one month of the number of daily confirmed cases of COVID-19 in Europe is ARIMA $(2, 1, 2)x(0, 0, 1)_7$, the AIC score of it is 2748.47. The ARIMA model of forecasting the future one month of the number of daily confirmed deaths of COVID-19 in the United States is ARIMA $(1, 0, 0)x(1, 0, 1)_7$, the AIC score of it is 1266.85. The ARIMA model of forecasting the future one month of the number of daily confirmed deaths of COVID-19 in Europe is ARIMA $(1, 0, 1)x(1, 0, 1)_7$, the AIC score of it is 1429.13. We used the function predict() in R language to forecast the future 30 days of the number of daily confirmed cases and deaths of COVID-19 in the

United States and Europe. Figure 5 shows the plot of forecasting the future 30 days of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe with 95% confidence intervals.

| | ARIMA model |
|---|---|
| Daily confirmed cases in US | $y_t = e_t - 0.54e_{t-1} + 0.33e_{t-7}$ |
| Daily confirmed cases in Europe | $y_t = 2154.12 + 0.67y_{t-1} + e_t + 0.93y_{t-7} - 0.46e_{t-7}$ |
| Daily confirmed deaths in US | $y_t = 1.11y_{t-1} - 0.72y_{t-2} + e_t - 1.49e_{t-1} + 0.76e_{t-2} + 0.41e_{t-7}$ |
| Daily confirmed deaths in Europe | $y_t = 4188.91 + 0.56y_{t-1} + e_t + 0.46e_{t-1} + 0.96y_{t-7} - 0.49e_{t-7}$ |

Here we can see that the equations of the best fitted ARIMA models for daily confirmed cases and deaths of the United States and Europe.

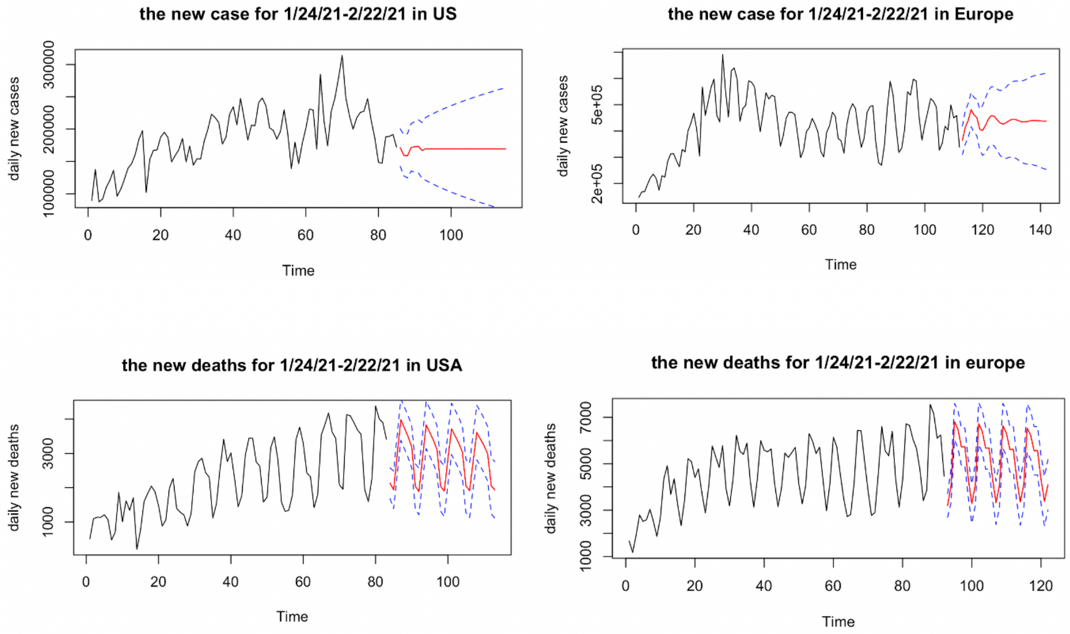


**Figure 5**. The plot of forecasting the future one month of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe.

From the first plot, we can see that the number of daily confirmed cases of COVID-19 in the United States initially increase for few days and then become stable in the rest of the month. In the second plot, the number of daily confirmed cases of COVID-19 in Europe initially decrease the amplitude of fluctuations and then finally becomes stable. In the third plot, the number of daily confirmed deaths of COVID-19 in the United States fluctuate as a seasonal pattern but the amplitude of fluctuations decreases. In the last plot, the number of daily confirmed deaths of COVID-19 in Europe also fluctuate as a seasonal pattern but the amplitude of fluctuations decreases. There are four tables that depict the forecasted values of daily confirmed cases and deaths of COVID-19 in the United States and Europe with 95% confidence intervals.

| DATE | LOWER CI | FORECAST | UPPER CI |
|---|---|---|---|
| 1/24/21 | 142238.8 | 171188.2 | 200137.7 |
| 1/25/21 | 127640.94 | 159392.4 | 191143.8 |
| 1/26/21 | 124448.86 | 158774.3 | 193099.7 |
| 1/27/21 | 135009.06 | 171728.5 | 208448 |
| 1/28/21 | 133711.93 | 172678.6 | 211645.3 |
| 1/29/21 | 132255.33 | 173346.5 | 214437.7 |
| 1/30/21 | 123870.37 | 166981.5 | 210092.6 |
| 1/31/21 | 120778.2 | 169424 | 218069.7 |
| 2/1/21 | 117787.35 | 169424 | 221060.6 |
| 2/2/21 | 114960.49 | 169424 | 223887.4 |
| 2/3/21 | 112273.29 | 169424 | 226574.6 |
| 2/4/21 | 109706.88 | 169424 | 229141 |
| 2/5/21 | 107246.32 | 169424 | 231601.6 |
| 2/6/21 | 104879.49 | 169424 | 233968.4 |
| 2/7/21 | 102596.43 | 169424 | 236251.5 |
| 2/8/21 | 100388.84 | 169424 | 238459.1 |
| 2/9/21 | 98249.68 | 169424 | 240598.2 |
| 2/10/21 | 96172.97 | 169424 | 242674.9 |
| 2/11/21 | 94153.53 | 169424 | 244694.4 |
| 2/12/21 | 92186.88 | 169424 | 246661 |
| 2/13/21 | 90269.07 | 169424 | 248578.8 |
| 2/14/21 | 88396.64 | 169424 | 250451.3 |
| 2/15/21 | 86566.51 | 169424 | 252281.4 |
| 2/16/21 | 84775.95 | 169424 | 254072 |
| 2/17/21 | 83022.48 | 169424 | 255825.4 |
| 2/18/21 | 81303.9 | 169424 | 257544 |
| 2/19/21 | 79618.2 | 169424 | 259229.7 |
| 2/20/21 | 77963.56 | 169424 | 260884.4 |
| 2/21/21 | 76338.33 | 169424 | 262509.6 |
| 2/22/21 | 74740.99 | 169424 | 264106.9 |

**Table 3**. The forecasted values of daily confirmed cases of COVID-19 in the United States with 95% confidence intervals.

| DATE | LOWER CI | FORECAST | UPPER CI |
|---|---|---|---|
| 1/24/21 | 309847.2 | 363964.1 | 418081.1 |
| 1/25/21 | 350224.3 | 413898.4 | 477572.5 |
| 1/26/21 | 374505.2 | 439428.9 | 504352.5 |
| 1/27/21 | 416240 | 481305.5 | 546370.9 |
| 1/28/21 | 395669.7 | 461515 | 527360.2 |
| 1/29/21 | 381702.5 | 451349.1 | 520995.6 |
| 1/30/21 | 329151.4 | 406026.9 | 482902.4 |
| 1/31/21 | 306708 | 402113.2 | 497518.3 |
| 2/1/21 | 316602.7 | 421256.3 | 525909.8 |
| 2/2/21 | 337309.1 | 445503 | 553696.9 |
| 2/3/21 | 348686.5 | 458688 | 568689.4 |
| 2/4/21 | 343477.8 | 455795.3 | 568112.7 |
| 2/5/21 | 326355.4 | 442973.2 | 559591 |
| 2/6/21 | 307767.5 | 430742.7 | 553717.8 |
| 2/7/21 | 296680.4 | 426394.1 | 556107.9 |
| 2/8/21 | 295205 | 430427 | 565649.1 |
| 2/9/21 | 298871.6 | 438097.8 | 577324 |
| 2/10/21 | 301375.1 | 443740.4 | 586105.7 |
| 2/11/21 | 299063.9 | 444470.2 | 589876.4 |
| 2/12/21 | 292317.9 | 441182.6 | 590047.2 |
| 2/13/21 | 284143.1 | 436976.6 | 589810.1 |
| 2/14/21 | 277663.7 | 434665.4 | 591667.2 |
| 2/15/21 | 274183.3 | 435140.3 | 596097.3 |
| 2/16/21 | 272850.7 | 437351.9 | 601853.1 |
| 2/17/21 | 271772.6 | 439479 | 607185.4 |
| 2/18/21 | 269472.7 | 440248.6 | 611024.6 |
| 2/19/21 | 265673.1 | 439562.2 | 613451.2 |
| 2/20/21 | 261123.7 | 438235.1 | 615346.4 |
| 2/21/21 | 256862.2 | 437250.7 | 617639.2 |
| 2/22/21 | 253502.3 | 437115.3 | 620728.3 |

**Table 4**. The forecasted values of daily confirmed cases of COVID-19 in Europe with 95% confidence intervals.

| DATE | LOWER CI | FORECAST | UPPER CI |
|---|---|---|---|
| 1/24/21 | 1681.148 | 2132.628 | 2584.108 |
| 1/25/21 | 1391.096 | 1936.697 | 2482.298 |
| 1/26/21 | 2396.981 | 2980.838 | 3564.694 |
| 1/27/21 | 3381.679 | 3982.33 | 4582.981 |
| 1/28/21 | 3128.095 | 3736.323 | 4344.55 |
| 1/29/21 | 2876.31 | 3487.994 | 4099.678 |
| 1/30/21 | 2575.816 | 3189.084 | 3802.353 |
| 1/31/21 | 1380.865 | 2039.005 | 2697.146 |
| 2/1/21 | 1209.023 | 1886.825 | 2564.626 |
| 2/2/21 | 2194.754 | 2881.419 | 3568.084 |
| 2/3/21 | 3138.863 | 3829.57 | 4520.277 |
| 2/4/21 | 2917.109 | 3609.669 | 4302.229 |
| 2/5/21 | 2691.123 | 3384.534 | 4077.946 |
| 2/6/21 | 2416.337 | 3110.141 | 3803.944 |
| 2/7/21 | 1315.365 | 2040.459 | 2765.552 |
| 2/8/21 | 1161.491 | 1900.546 | 2639.6 |
| 2/9/21 | 2084.171 | 2829.565 | 3574.959 |
| 2/10/21 | 2966.53 | 3714.825 | 4463.12 |
| 2/11/21 | 2760.732 | 3510.358 | 4259.985 |
| 2/12/21 | 2550.568 | 3300.807 | 4051.046 |
| 2/13/21 | 2294.654 | 3045.175 | 3795.695 |
| 2/14/21 | 1272.071 | 2047.693 | 2823.315 |
| 2/15/21 | 1130.422 | 1917.332 | 2704.241 |
| 2/16/21 | 1991.855 | 2783.907 | 3575.96 |
| 2/17/21 | 2815.231 | 3609.64 | 4404.049 |
| 2/18/21 | 2623.487 | 3418.979 | 4214.47 |
| 2/19/21 | 2427.572 | 3223.562 | 4019.551 |
| 2/20/21 | 2188.937 | 2985.155 | 3781.374 |
| 2/21/21 | 1237.964 | 2054.83 | 2871.695 |
| 2/22/21 | 1107.054 | 1933.252 | 2759.45 |

**Table 5**. The forecasted values of daily confirmed deaths of COVID-19 in the United States with 95% confidence intervals.

| DATE | LOWER CI | FORECAST | UPPER CI |
|---|---|---|---|
| 1/24/21 | 2692.316 | 3199.09 | 3705.864 |
| 1/25/21 | 3322.978 | 4051.056 | 4779.134 |
| 1/26/21 | 6029.762 | 6816.199 | 7602.636 |
| 1/27/21 | 5678.568 | 6482.974 | 7287.381 |
| 1/28/21 | 4916.233 | 5726.366 | 6536.499 |
| 1/29/21 | 4920.492 | 5732.468 | 6544.445 |
| 1/30/21 | 3526.715 | 4339.286 | 5151.858 |
| 1/31/21 | 2430.626 | 3281.318 | 4132.01 |
| 2/1/21 | 3194.432 | 4081.539 | 4968.646 |
| 2/2/21 | 5831.335 | 6729.903 | 7628.472 |
| 2/3/21 | 5501.015 | 6403.26 | 7305.504 |
| 2/4/21 | 4768.685 | 5672.115 | 6575.545 |
| 2/5/21 | 4772.171 | 5675.984 | 6579.798 |
| 2/6/21 | 3431.089 | 4335.026 | 5238.963 |
| 2/7/21 | 2383.284 | 3316.933 | 4250.581 |
| 2/8/21 | 3122.855 | 4086.136 | 5049.418 |
| 2/9/21 | 5660.183 | 6632.856 | 7605.528 |
| 2/10/21 | 5342.914 | 6318.604 | 7294.294 |
| 2/11/21 | 4638.731 | 5615.395 | 6592.06 |
| 2/12/21 | 4642.099 | 5619.078 | 6596.058 |
| 2/13/21 | 3352.377 | 4329.459 | 5306.54 |
| 2/14/21 | 2347.796 | 3350.347 | 4352.898 |
| 2/15/21 | 3061.951 | 4090.082 | 5118.213 |
| 2/16/21 | 5502.985 | 6539.254 | 7575.524 |
| 2/17/21 | 5198.149 | 6237.037 | 7275.925 |
| 2/18/21 | 4521.027 | 5560.761 | 6600.494 |
| 2/19/21 | 4524.295 | 5564.302 | 6604.309 |
| 2/20/21 | 3283.98 | 4324.076 | 5364.171 |
| 2/21/21 | 2320.189 | 3382.465 | 4444.741 |
| 2/22/21 | 3009.214 | 4093.867 | 5178.52 |

**Table 6**. The forecasted values of daily confirmed deaths of COVID-19 in Europe with 95% confidence intervals.

These four tables show us the result of forecasting the daily confirmed cases and daily confirmed deaths of COVID-19 in the United States and Europe with 95% confidence intervals. Hence, in order to check the performance of the ARIMA models, a comparison between the forecasted curve of future one month of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe and the actual curve of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe is shown by the Figure 6 to Figure 9.

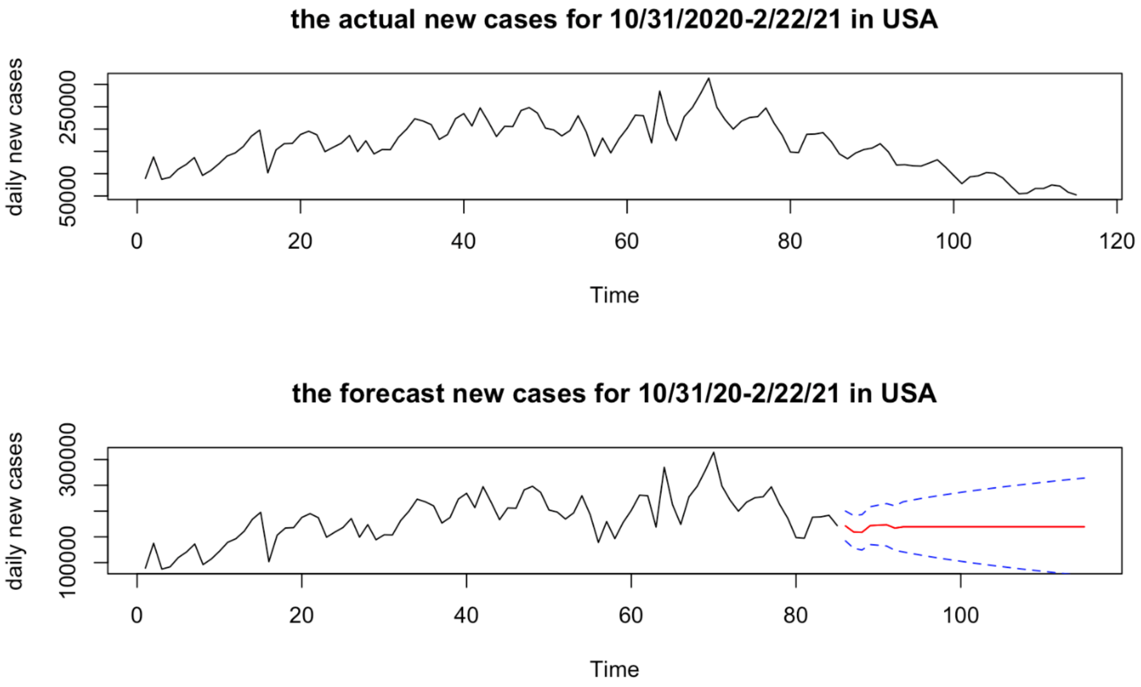


**Figure 6**. The comparison between the forecasted curve of daily confirmed cases of COVID-19 in the US with the actual curve of daily confirmed cases of COVID-19 in the US.

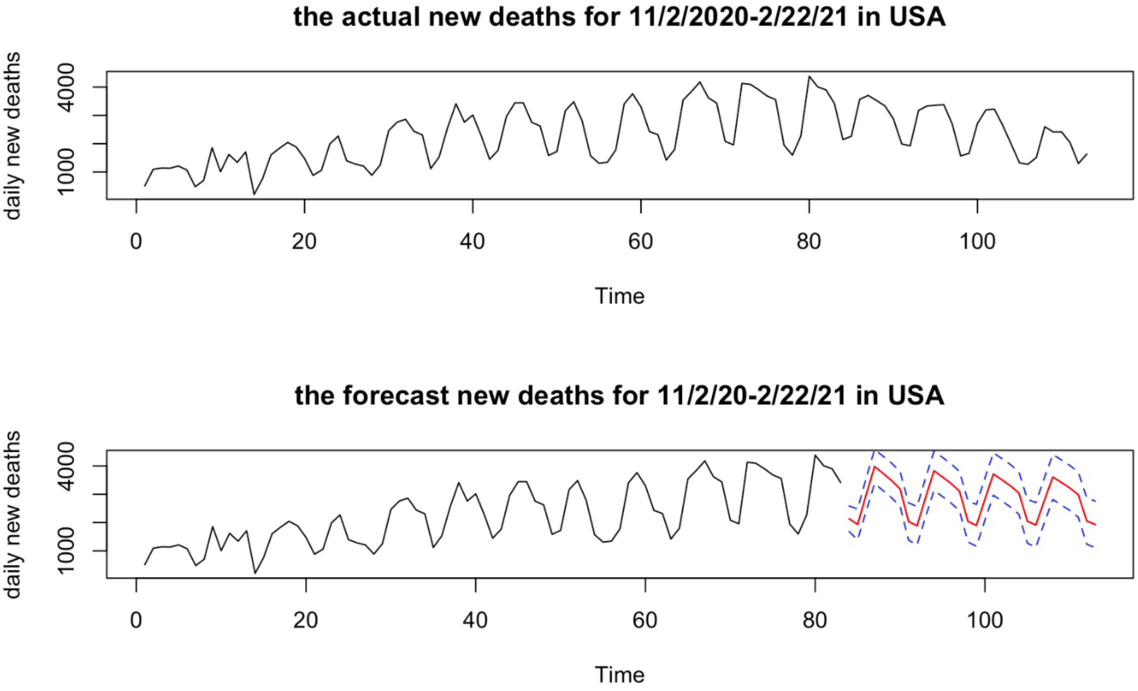
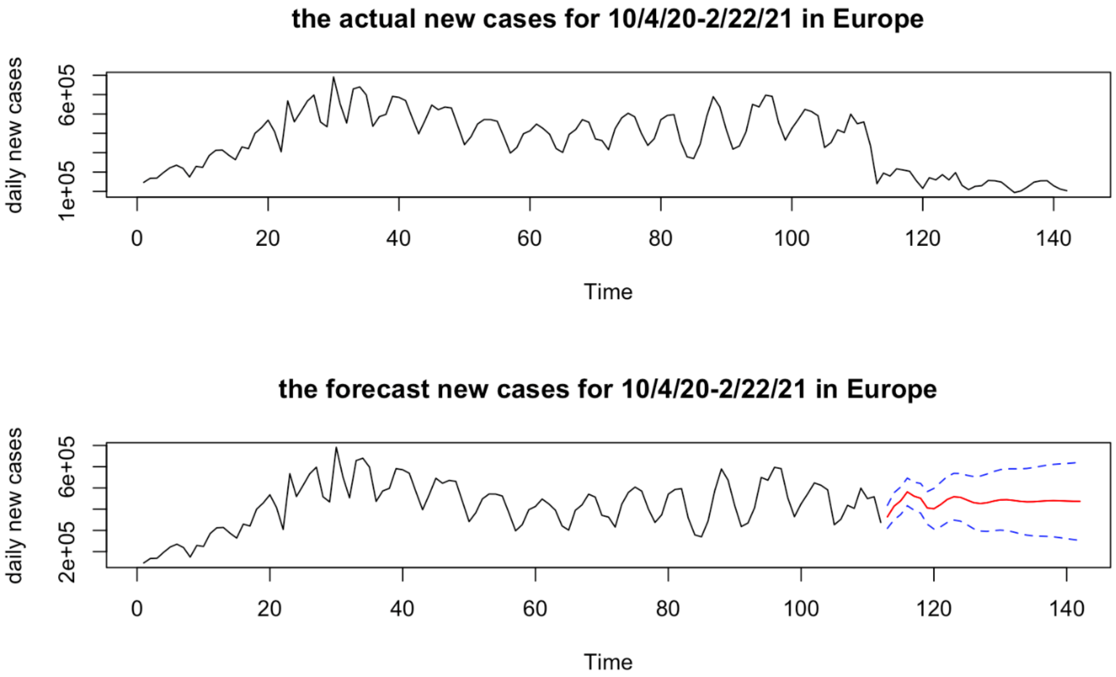


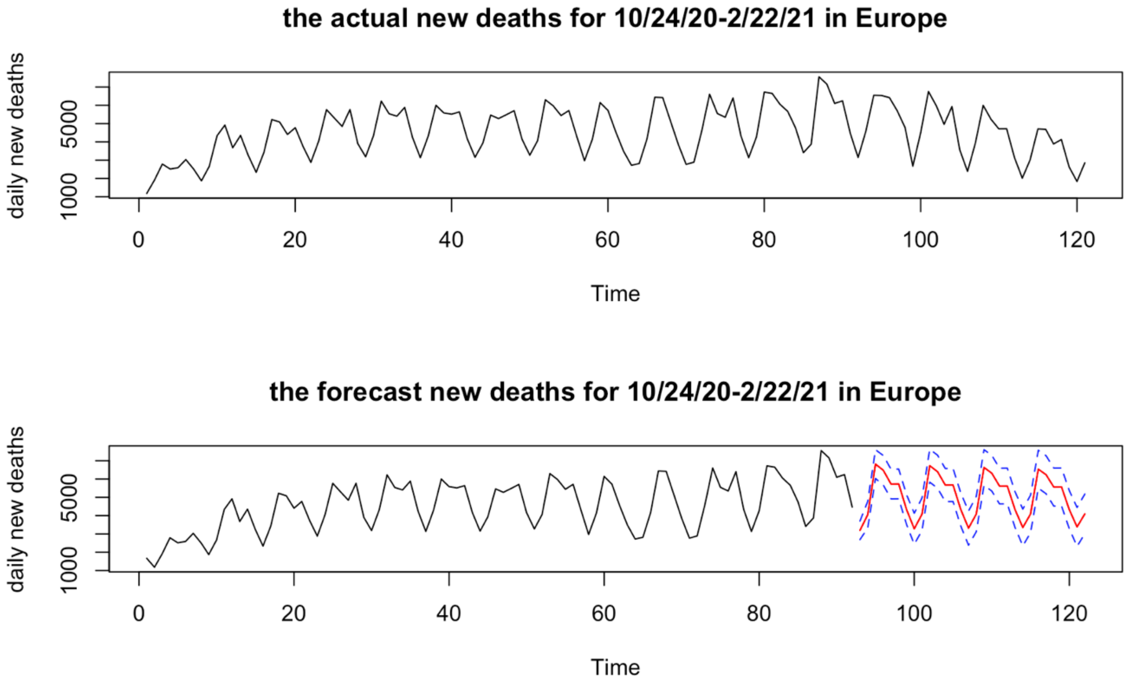
**Figure 7**. The comparison between the forecasted curve of daily confirmed deaths of COVID-19 in the US with the actual curve of daily confirmed deaths of COVID-19 in the US.



the actual new cases for 10/4/20-2/22/21 in Europe

the forecast new cases for 10/4/20-2/22/21 in Europe

**Figure 8**. The comparison between the forecasted curve of daily confirmed cases of COVID-19 in Europe with the actual curve of daily confirmed cases of COVID-19 in Europe.


the actual new deaths for 10/24/20-2/22/21 in Europe

the forecast new deaths for 10/24/20-2/22/21 in Europe

**Figure 9**. The comparison between the forecasted curve of daily confirmed deaths of COVID-19 in Europe with the actual curve of daily confirmed deaths of COVID-19 in Europe.

Figure 6 to Figure 9 show us the comparison between the forecasted curve of future 30 days of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe and the actual curve of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe. We can see that all the actual curves are within the confidence interval, but the trend is off in certain cases. This could be due to the strict control measures of the government and the increase in the number of vaccinations. In the following Table 7, the comparison between the forecasted value of future one month of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe and the actual value of the number of daily confirmed cases and deaths of COVID-19 in the United States and Europe is shown.

| DATE | UNITED STATES | | | | EUROPE | | | |
|---|---|---|---|---|---|---|---|---|
| | Daily new cases | | Daily new deaths | | Daily new cases | | Daily new deaths | |
| | ACTUAL | FORECAST | ACTUAL | FORECAST | ACTUAL | FORECAST | ACTUAL | FORECAST |
| 1/24/21 | 144234 | 171188.2 | 2147 | 2132.628 | 139630 | 363964.1 | 3144 | 3199.090 |
| 1/25/21 | 133454 | 159392.4 | 2261 | 1936.697 | 194468 | 413898.4 | 4594 | 4051.056 |
| 1/26/21 | 146448 | 158774.3 | 3567 | 2980.838 | 179078 | 439428.9 | 6549 | 6816.199 |
| 1/27/21 | 153985 | 171728.5 | 3706 | 3982.33 | 216411 | 481305.5 | 6535 | 6482.974 |
| 1/28/21 | 157306 | 172678.6 | 3525 | 3736.323 | 210481 | 461515 | 6421 | 5726.366 |
| 1/29/21 | 167111 | 173346.5 | 3335 | 3487.994 | 203998 | 451349.1 | 5704 | 5732.468 |
| 1/30/21 | 148824 | 166981.5 | 2879 | 3189.084 | 155901 | 406026.9 | 4788 | 4399.286 |
| 1/31/21 | 119367 | 169424 | 1991 | 2039.005 | 115693 | 402113.2 | 2669 | 3281.318 |
| 2/1/21 | 120200 | 169424 | 1925 | 1886.825 | 170705 | 421256.3 | 4497 | 4081.539 |
| 2/2/21 | 117616 | 169424 | 3180 | 2881.419 | 159485 | 445503 | 6755 | 6729.903 |
| 2/3/21 | 116960 | 169424 | 3330 | 3829.57 | 186316 | 458688 | 5976 | 6403.260 |
| 2/4/21 | 123907 | 169424 | 3359 | 3609.669 | 159317 | 455795.3 | 4959 | 5672.155 |
| 2/5/21 | 131146 | 169424 | 3379 | 3384.534 | 196844 | 442973.2 | 5929 | 5675.984 |
| 2/6/21 | 114557 | 169424 | 2694 | 3110.141 | 130722 | 430742.7 | 3567 | 4335.026 |
| 2/7/21 | 95994 | 169424 | 1569 | 2040.459 | 109040 | 426394.1 | 2383 | 3316.933 |
| 2/8/21 | 77737 | 169424 | 1659 | 1900.546 | 125972 | 430427 | 3936 | 4086.136 |
| 2/9/21 | 92986 | 169424 | 2715 | 2829.565 | 129190 | 438097.8 | 5994 | 6632.856 |
| 2/10/21 | 95194 | 169424 | 3193 | 3714.825 | 156265 | 443740.4 | 5234 | 6318.604 |
| 2/11/21 | 102417 | 169424 | 3223 | 3510.358 | 154852 | 444470.2 | 4713 | 5615.395 |
| 2/12/21 | 101030 | 169424 | 2651 | 3300.807 | 148303 | 441182.6 | 4716 | 5619.078 |
| 2/13/21 | 90642 | 169424 | 1990 | 3045.175 | 121070 | 436976.6 | 3145 | 4329.459 |
| 2/14/21 | 72164 | 169424 | 1321 | 2047.693 | 93956 | 434665.4 | 2014 | 3350.347 |
| 2/15/21 | 55077 | 169424 | 1272 | 1917.332 | 102066 | 435140.3 | 3013 | 4090.082 |
| 2/16/21 | 56312 | 169424 | 1512 | 2783.907 | 123071 | 437351.9 | 4708 | 6539.254 |
| 2/17/21 | 66839 | 169424 | 2600 | 3609.64 | 148007 | 439479 | 4685 | 6237.037 |
| 2/18/21 | 66824 | 169424 | 2415 | 3418.979 | 154514 | 440248.6 | 3887 | 5560.761 |
| 2/19/21 | 74676 | 169424 | 2417 | 3223.562 | 155066 | 439562.2 | 4128 | 5564.302 |
| 2/20/21 | 72354 | 169424 | 2040 | 2985.155 | 128859 | 438235.1 | 2621 | 4324.076 |
| 2/21/21 | 58702 | 169424 | 1300 | 2054.83 | 111977 | 437250.7 | 1836 | 3382.465 |
| 2/22/21 | 52530 | 169424 | 1635 | 1933.252 | 103446 | 437115.3 | 2857 | 4093.867 |

**Table 7**. The comparison between the forecast value of daily confirmed cases and deaths of COVID-19 in the US and Europe with the actual value of daily confirmed cases and deaths of COVID-19 in the US and Europe.

## DISCUSSION

From the results we obtained in the last section, we can conclude that considering the change point of data can help us to fit a good prediction ARIMA model. According to the Figure 3 and Figure 4, the normal probability plots of the residuals of the ARIMA models become linear after we consider the change point of the data and thus validates the fitted models. This obviously prove that setting change points of the data is helpful for fitting a good prediction ARIMA model.

In addition, from the Figure 2, we can see that for both Europe and US, the change point for deaths was more than the change point of the new cases. For instance, in the United States the change point for the new confirmed cases was No.92 but the change point for the new deaths was No. 94. In Europe, it is same as the United States, the change point for the new confirmed cases was No.65 but the change point for the new deaths was No. 85. This proves that the death cases follow the number of infected cases. The difference between the US and Europe is probably because of the speed of deaths is slower than the speed of increasing new confirmed cases.

From the Table 7, both the actual value of confirmed cases and confirmed deaths are smaller than the forecasting value. This is probably because the intervention of vaccination and the strict control measures of public services. It is a good result for health system and hospital since the services staff do not need to prepare more facilities for new patients.

# CONCLUSION

Based on the results we obtained, we can make a conclusion that finding change points of data when building ARIMA models is necessary for the models forecasting. By checking the residuals plot of the ARIMA models, we can see that the ARIMA model fitted by new data is better than the ARIMA model fitted by the original data. However, the data after Jan 2021 shows actual values below forecasting values. This may be due to intervention of vaccination. Therefore, we need another change point detection since the ARIMA model might have shifted. We can find new change point when we collect enough data, and thus a new prediction ARIMA model with an additional change point would be justifiable.

# BIBLIOGRAPHY

1. https://otexts.com/fpp2/arima.html

2. https://covid.cdc.gov/covid-data-tracker/#trends_dailytrendscases

3. https://ourworldindata.org/covid-cases