Short-term Forecast of Confirmed COVID-19 Infected Cases

Li Sun
Department of Mathematics and Statistics, Georgia State University, Atlanta, 30303, USA

Abstract: The outbreak of the COVID-19 was identified in Wuhan, China, in December 2019. From then on, it has rapidly spread to almost all countries on the globe, and was recognized as a pandemic by the World Health Organization on March 11th 2020. As the number of coronavirus cases reported rapidly increases, the spread of coronavirus is a serious threat to global health. In this work, the total confirmed COVID19 cases in the countries and the USA states are forecasted up until March 22, 2020. Fitting time series analysis and statistical algorithms are implemented to produce the best short term prediction. An online Kalman filter, which runs in real-time, provides a very good one-day prediction for countries and USA states, and time series ARIMA models generate impressive 9-day advance predictions for the USA and Italy.

1. Introduction

The ongoing epidemic of COVID-19 began in Hubei Province, China, in December 2019 and continues to cause infections all over the globe. On March 11th, the World Health Organization declared the rapidly spreading coronavirus outbreak a pandemic. In the USA, President Donald Trump declared a national emergency to help handle the growing outbreak of COVID-19 on March 13[1]. As the numbers of cases and deaths continue to accumulate every day, almost all the countries around the world are struggling to contain the spread of the virus. Even with strict lockdowns combined with isolation and quarantine measures, the trend of transmission is still unclear.

Because the COVID-19 epidemiological data is limited, short-term forecasts can be useful to guide the allocation of resources critical to bringing the epidemic under control [2]. In this work, online Kalman filter and time series ARIMA models are used to generate 1-day and 9-day ahead forecasts of cumulative confirmed cases in countries of the Global and states in the USA.

2. Methods

2.1 Data

Data of total confirmed COVID19 cases was obtained from Github for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The time-series data was derived from the daily case reports and was updated daily. However, aiming to provide a cleaner and more organized dataset, the Johns Hopkins Data Source updated the time series tables after March 22. Due to this reason, the dataset used in this work only contains the posted dashboard case reports from Jan 22 to March 22 for COVID-19.

The dataset contains 501 observations and 66 attributes.

2.1.1 Field description

There are 66 attributes in the dataset. The first attribute is Province/State, the second one is Country/Region, the third and fourth are Lat and Long coordinates, and the others are dates from January 22 to March 22. Their brief descriptions are listed as follows:

- Province/State:
    - China - province name
    - US/Canada/Australia/ - city name, state/province name
    - Others - the name of the event (e.g., "Diamond Princess" cruise ship)
    - other countries - blank
- Country/Region:
    - country/region name conforming to the WHO
- Lat and Long:
    - a coordinates corresponding to each Country/Region and Province/State
- 1/22/20 -- 3/22/20
    - Number of confirmed cases on each date, which has the format of MM/DD/YY

2.1.2 Import Data

The data is read online from Github, which contains the daily total cases of confirmed per location according to the following data of the world health organization.

To keep consistent with WHO, ‘Mainland China’ and ‘US’ in the Country/Region are replaced with ‘China’ and ‘United States’ respectively.

2.1.3 Data Aggregation and Separation

With the aim of integrity and organizational clarity, the original data is separated into Global data and USA data. One dataset is global data, which contains the records for all countries with confirmed COVID-19 cases. The other one is the USA data, which contains all USA-related records.

However, in the original dataset, there are 8 countries whose data were reported inconsistently. Their records were split into provinces/states, even counties in the USA, at beginning dates with their countries' record blank. Then the records went back to under countries and kept the province/state/county blank. There is even no record for the United States as a country. In order to resolve this issue, all records in each date are aggregated from each country's province/state to form a completed record for each date of each country. At the same time, all USA related data are extracted out from the original dataset and a new dataset called the USA is generated.

After aggregation and separation, there are 182 observations in the Global dataset. This means all 182 countries in the whole world have detected COVID-19 on March 22, 2020. There are 56 observations in the USA dataset. All 50 states, Washington D.C, and all of their 3 territories (Guam, Puerto Rico and US Virgin Islands) have confirmed cases. The other two are from Grand Princess and Diamond Princess Cruises.

2.1.4 Data Preprocessing

In the Global dataset, the 8 countries mentioned above have no *Lat* and *Long* values. Python's country information package is called to retrieve the capitals of these countries. 'Geocoders' is imported from the 'geopy' package, and its function 'Geonames' is operated to obtain the geometric location of longitude and latitude coordinates of these countries.

2.1.5 Dataset Extraction

The data for the countries of the USA and Italy are extracted from the Global data set as two separate time series. Two ARIMA models are constructed based on these two-time series, and they provide us 9-day forecasts on confirmed cases for these two most severe countries.

2.2 Data Visualization

The data visualization technique communicates insights from data through visual representation. By refining large datasets into visual graphics, viewers can easily understand the complex relationships within the data.

'Plotly' in python allows interactive features in visual plots. In this work, several packages are imported for different situations. The plots are shown to present time series interactive visualization about confirmed COVID-19 cases in the countries/states. The maps are produced to show the confirmed cases among different locations.

2.3 Algorithm

2.3.1 Kalman filter

The Kalman filter is a widely applied concept in time series analysis. It is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables [2].

The Kalman filter is a recursive algorithm. It first produces estimates of the current state variables, and then a weighted average is used to update these estimates with more weight assigned to the higher certainty. The appendix includes a detailed description of the model and its parameters.

In this work, the Kalman filter is running online in real-time. Each day, new observations as the present input measurements, along with the previously calculated state and its uncertainty matrix, are used to update the algorithm. Then the algorithm does the estimation and prediction for the next day. All the processes are done online.

2.3.2 Time Series ARIMA

A time series is a data sequence ordered by time. It is discrete, and the interval between each point is constant [4]. The ARIMA model is a popular and widely used statistical method for time series forecasting.

ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. It is a class of models that captures a suite of different standard temporal structures in time

series data [5]. An ARIMA model includes three parameters. They are model lag order p, degree of differencing d and the order of moving average q.

In this work, 9-day forecasts are generated on the time series of confirmed COVID-19 cases in the USA and Italy. Before ARIMA models are established, the stationarities of these two-time series are tested. A log transform is applied on USA data to guarantee the data stationarity.

2.4 Short-term forecast

An online Kalman filter algorithm is implemented to fit the global and USA data, and a prediction for one day ahead is generated for each country and state. Because the reported data is available beginning January 22 and archived on March 22, 2020, the 1-day prediction period includes daily prediction in these two months range. The basic estimator parameters, such as mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) between the model predictions and data are calculated for each country and state. Meanwhile, the comparison graphs between predicted values and real data are plotted. Since it is an online algorithm and each prediction summarizes the previous online predictions, there is no overfitting or bias [6].

Time Series ARIMA models are conducted to achieve 9-day forecasts until the end of March for two countries: the USA and Italy. As mentioned above, the sample dates are from January 22 to March 22, so that the total sample number for each country is 60. Before the models are built, the Rolling Statistics and Augmented Dickey-Fuller Test are used to determine whether the given time series is stationary. The best parameters for each model are chosen according to ACF/PACF plots, AIC and variance values. After models are established and model's residuals are checked, the models are refitted back to the original time series for comparison, following with conduction of 9-day predictions.

3. Implementation and Results

3.1 Data Visualization for Original Datasets

The major countries in global and major states in the USA with high confirmed COVID-19 cases are visually presented. Figures 1 and 2 contain the confirmed case for the top 10 countries and USA states from January 22 - March 22, 2020, respectively.

3.1.1 Top 10 countries within Global data

Figure 1 shows the top10 countries with the highest confirmed cases (figure 1).

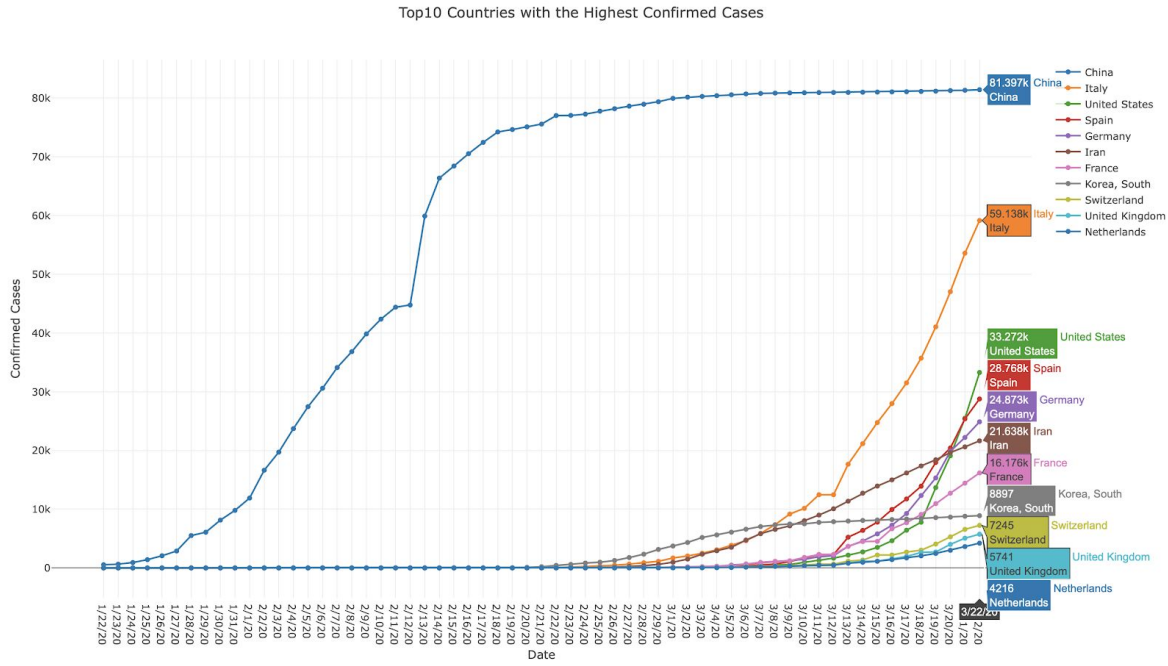


Figure 1: The top 10 countries with the highest confirmed cases by March 22, 2020

As we can see there is a high jump in the China curve (blue line) around February 12-14, which indicates a clear eruption trend in China at that time. Then, COVID-19 was transmitted into the other countries and rapidly spread all over the world around the end of February. In the middle of March, the upward trends of European countries and the USA became sharp. Their situations worsened from then on. Meanwhile, the trends in China and South Korea (gray line) plateaued , which points out a major decrease in COVID-19 spread. The Chinese Government's lockdown policy, and South Korea's expansive diagnostic capacity at scale and extensive contact tracing successfully prevented the virus from spreading.

3.1.2 Time series in USA data

Figure 2 shows the top10 states with the highest confirmed cases (figure 2).

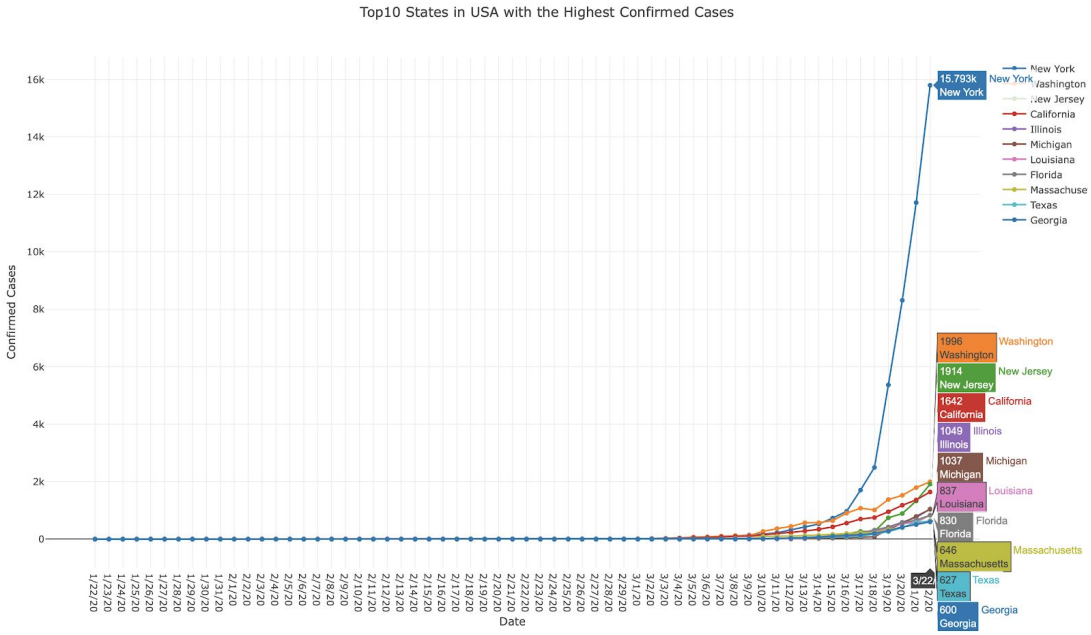


Figure 2: The top 10 USA states with the highest confirmed cases by March 22, 2020.

It is stunning to notice that the trend in New York. It is so sharp that it implies the future of New York is going to be very tough. There are nearly 16K confirmed cases in New York by March 22. Like Hubei province, New York became the epidemic center of the USA. The following two states are Washington and California. Both of them have confirmed about 2000 cases.

3.2 Kalman Filter for 1-day Forecast

The online Kalman algorithm is written in R script. 'Rpy2' package is called to upload R script and make R script executable inside python. The standard deviation of the noise vector in control-input, which is used to calculate the covariance Q of process noise vector, is set as 0.1. The covariance R of measurement noise vector is set 0.1 too. After parameter initialization and time difference adjustment, the model produces a 1-day forecast for each country and state. Statistics measurements, such as MSE, RMSE, and MAE are calculated for model evaluation. The script allows users to select regions and get their plots and predictions. Prediction and evaluation results for several samples are displayed below.

The visualizations out of the prediction are presented as well. Package '*scattergeo*' from '*plotly*' is implemented to create a scatter map for confirmed cases in Global. Package '*choropleth*' from '*plotly*' is implemented to generate a choropleth map for the states in the USA. Since the Kalman algorithm is executed online, the maps are updated for each day in a real-time manner.

3.2.1 Kalman Fitting Models

3.2.1.1 United States and Italy

In the Global data, the USA and Italy as two sample countries are presented. We can see they are an almost perfect fit. Kalman's predictions successfully follow their trend.

Figure 3 illustrated the prediction of fitness plots. The orange lines represent the Kalman prediction, and the blue lines are the actual values of confirmed cases. The statistics measurements are displayed in the yellow panels. The RMSE is 626.6 for the United States, and 373.2 for Italy. Table 1 gives a comparison between the actual values of confirmed cases and Kalman prediction values in the newest 10 dates. For example, In the United States, on 2020-03-22, Kalman predicted 31782 cases while there were 33272. For tomorrow (2020-03-23) Kalman predicts 40674, which means there should be 6947 (40674-33272=6947) new confirming cases detected tomorrow.

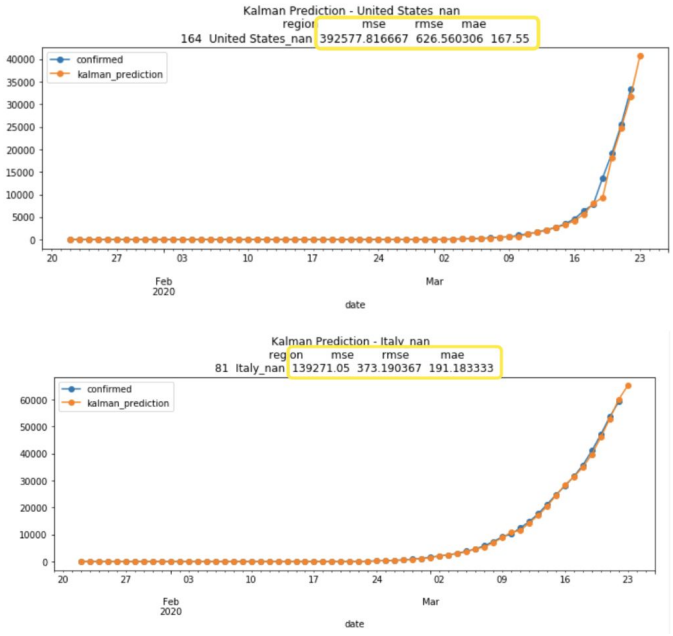


Figure 3: Prediction fitness plots for the USA (upper plot) and Italy (lower plot) countries. Their evaluation results are shown in the yellow panels.

date	region	confirmed	kalman_prediction
2020-03-14	United States_nan	2727.0	2656
2020-03-15	United States_nan	3499.0	3271
2020-03-16	United States_nan	4632.0	4208
2020-03-17	United States_nan	6421.0	5661
2020-03-18	United States_nan	7783.0	8025
2020-03-19	United States_nan	13677.0	9321
2020-03-20	United States_nan	19100.0	18160
2020-03-21	United States_nan	25489.0	24796
2020-03-22	United States_nan	33272.0	31782
2020-03-23	United States_nan	NaN	40674

date	region	confirmed	kalman_prediction
2020-03-14	Italy_nan	21157.0	20429
2020-03-15	Italy_nan	24747.0	24504
2020-03-16	Italy_nan	27980.0	28355
2020-03-17	Italy_nan	31506.0	31364
2020-03-18	Italy_nan	35713.0	34938
2020-03-19	Italy_nan	41035.0	39693
2020-03-20	Italy_nan	47021.0	46034
2020-03-21	Italy_nan	53578.0	52870
2020-03-22	Italy_nan	59138.0	60040
2020-03-23	Italy_nan	NaN	65076

Table 1: Comparison between Confirmed cases and Kalman prediction values in the newest 10 dates for the USA (left) and Italy (right)

From both the plot and table of the United States, there is a sharp positive trend. We can see the Kalman model mispredicted the big jump on 03-19, but it caught up with the actual values very quickly on 03-20.

3.2.1.2 New York State and Washington State

In the USA data, New York and Washington as two sample states are presented. Here we can see almost perfect predictions as well. Kalman's predictions successfully follow their trend.

Figure 4 graphs the prediction of fitness plots. The orange lines represent the Kalman prediction, and the blue lines are the actual values of confirmed cases. The statistics measurements are displayed in the yellow panels. The RMSE is 311.6 for New York, and 68.4 for Washington. Table 2 gives a comparison between the actual values of confirmed cases and Kalman prediction values in the newest 10 dates. In New York on 2020-03-22, Kalman predicted 15074 cases while there were 15793. For tomorrow (2020-03-23) Kalman predicts 19701, which means there should be 3908 (19701-15793=3908) new confirming cases detected on 2020-03-23. Recalling the new confirmed cases from Kalman's prediction for 2020-03-23 should be 6947, there will be more than 56% of new confirmed cases coming from New York.

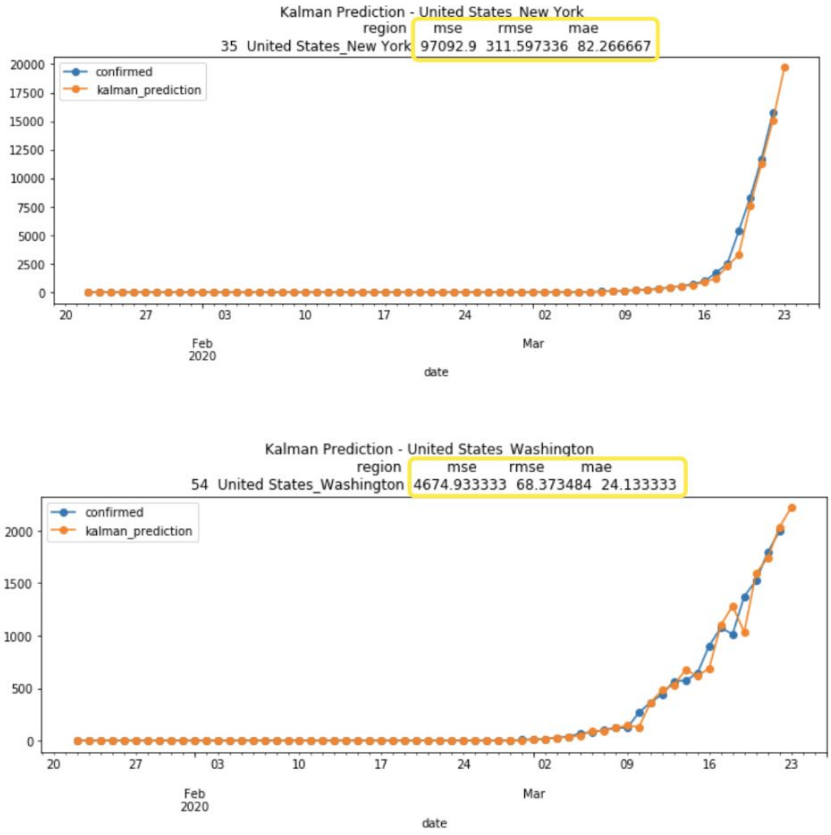


Figure 4: Prediction fitness plots for the New York state (upper) and Washington states (lower) in the USA. Their evaluation results are shown in the yellow panels.

date	region	confirmed	kalman_prediction
2020-03-14	United States_New York	525.0	521
2020-03-15	United States_New York	732.0	628
2020-03-16	United States_New York	967.0	907
2020-03-17	United States_New York	1706.0	1196
2020-03-18	United States_New York	2495.0	2292
2020-03-19	United States_New York	5365.0	3286
2020-03-20	United States_New York	8310.0	7603
2020-03-21	United States_New York	11710.0	11304
2020-03-22	United States_New York	15793.0	15074
2020-03-23	United States_New York	NaN	19701

date	region	confirmed	kalman_prediction
2020-03-14	United States_Washington	572.0	677
2020-03-15	United States_Washington	643.0	614
2020-03-16	United States_Washington	904.0	691
2020-03-17	United States_Washington	1076.0	1101
2020-03-18	United States_Washington	1014.0	1284
2020-03-19	United States_Washington	1376.0	1034
2020-03-20	United States_Washington	1524.0	1594
2020-03-21	United States_Washington	1793.0	1739
2020-03-22	United States_Washington	1996.0	2036
2020-03-23	United States_Washington	NaN	2219

Table 2: Comparison between Confirmed cases and Kalman prediction values in the newest 10 dates for the USA (upper plot) and Italy (lower plot) countries.

One more thing to mention here is the data in Washington state. The confirmed cases on 03-17 are higher than that on 03-18; there exist some doubts here. If the real meaning is ignored, we can see the Kalman model adapts this fluctuation very fast, and it works very well.

3.2.2 Prediction Visualization

3.2.2.1 Prediction map for the Global

A Scattergeo map is created showing the number of confirmed cases for global countries. It is updated on a daily basis. The map shows the entire world. Each country is located based on its *Lat* and *Long* coordinates in the Global dataset. The size of the marker is proportional to the number of confirmed cases per country. Its text message will hover while the mouse cursor is on a country. The information on location, country name, date, and numbers of confirmed and Kalman predicted hover in the text message panel (tooltip window).

Figure 5 is the Scattergeo map for the Global on March 22. China has the biggest circle, which indicates that China had the most cases at that time. There are big circles clustered in Europe. This signals a critical spread of the virus there, as Europe is now the epicenter of the COVID-19 pandemic.

The marker size of the USA is the third largest, portraying that the USA is N. 3 with the highest cases. The text message panel shows that there are 33272 confirmed cases and 31782 predicted cases in the USA on 2020-03-22. It flags a warning sign that the USA is on the same trajectory as countries in Europe.

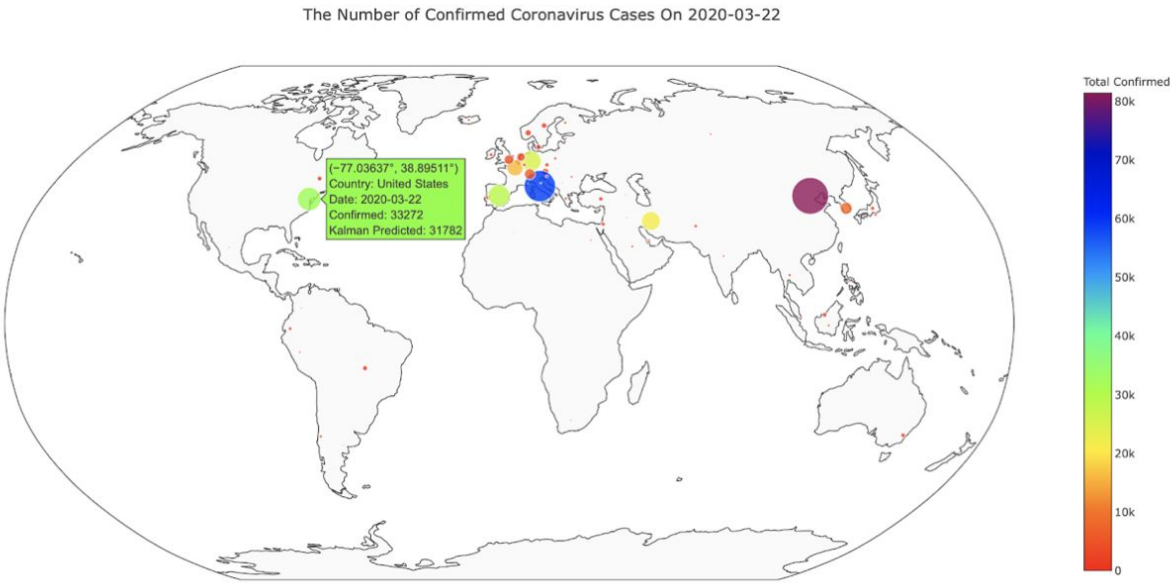


Figure 5: Scattergeo map for the Global on 2020-03-22.

3.2.2.2 Prediction map for the USA

A Choropleth map is created showing the number of confirmed cases for USA states. It is updated on a daily basis too. The map shows the entire USA. Each state is located based on its *Lat* and *Long* coordinates in the USA dataset. The color scale is used to represent the number of confirmed cases. The darker the color is, the more cases a state has. When the mouse cursor hovers over each state, the state name, date, and numbers of confirmed and Kalman predicted are displayed in the text message panel (tooltip window).

Figure 6 is the Choropleth map for the USA on 2020-03-22. New York has the darkest color, which shows that New York has the most cases at that time. And since all the other states have a very light color compared with New York, the confirmed cases in New York have a higher scale level (10X more) than others.

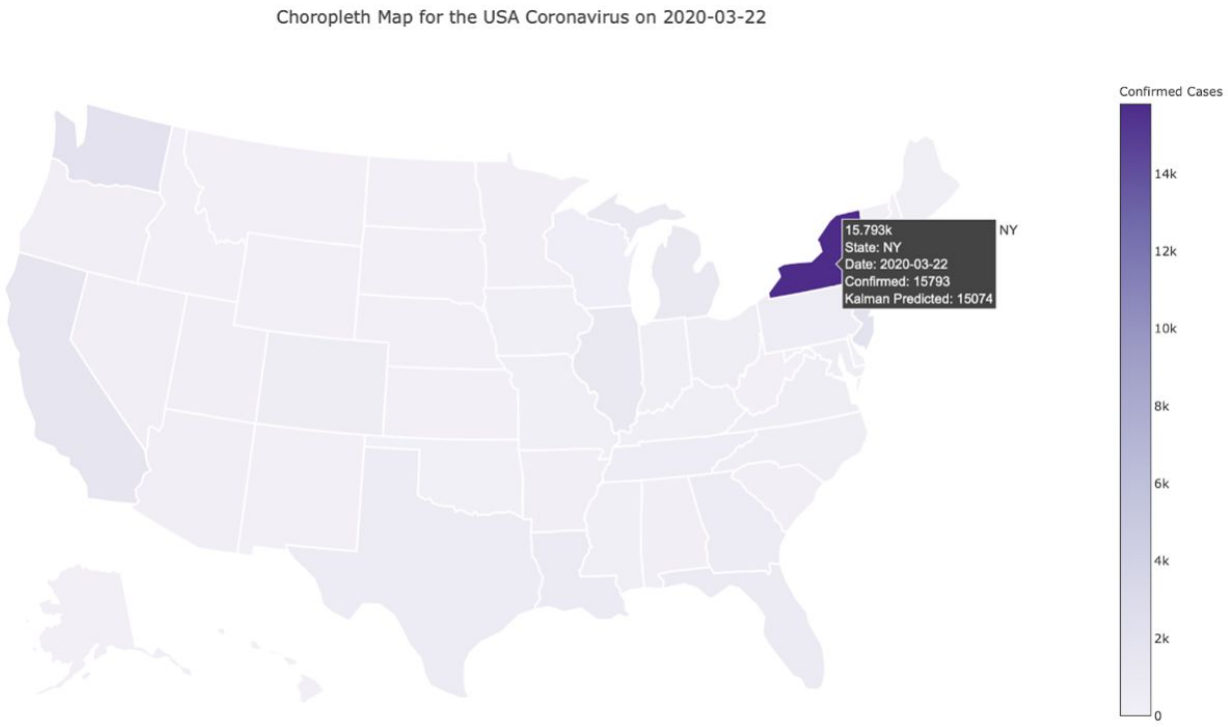


Figure 6: Choropleth map for the USA states on 2020-03-22

3.3 ARIMA Model for 9-day Forecast

Time Series ARIMA models are conducted to achieve 9-day forecasts, from March 23 to the end of March, which is March 31, for two countries: the USA and Italy. There are 60 sample data in each time series.

3.3.1 ARIMA model for the USA

3.3.1.1 Data visualization of the original USA data

A bar chart of the original data provides a glance at the relationship between the date and confirmed cases. The date range is from January 22 to March 22. There are a total of 60 observations in the dataset.

Figure 7 provides a bar chart of confirmed coronavirus cases in the USA. The graph presents the dates with rectangular bars and with heights proportional to the values of coronavirus cases. We can see the cases likely increase exponentially. The increasing trend is sharp.

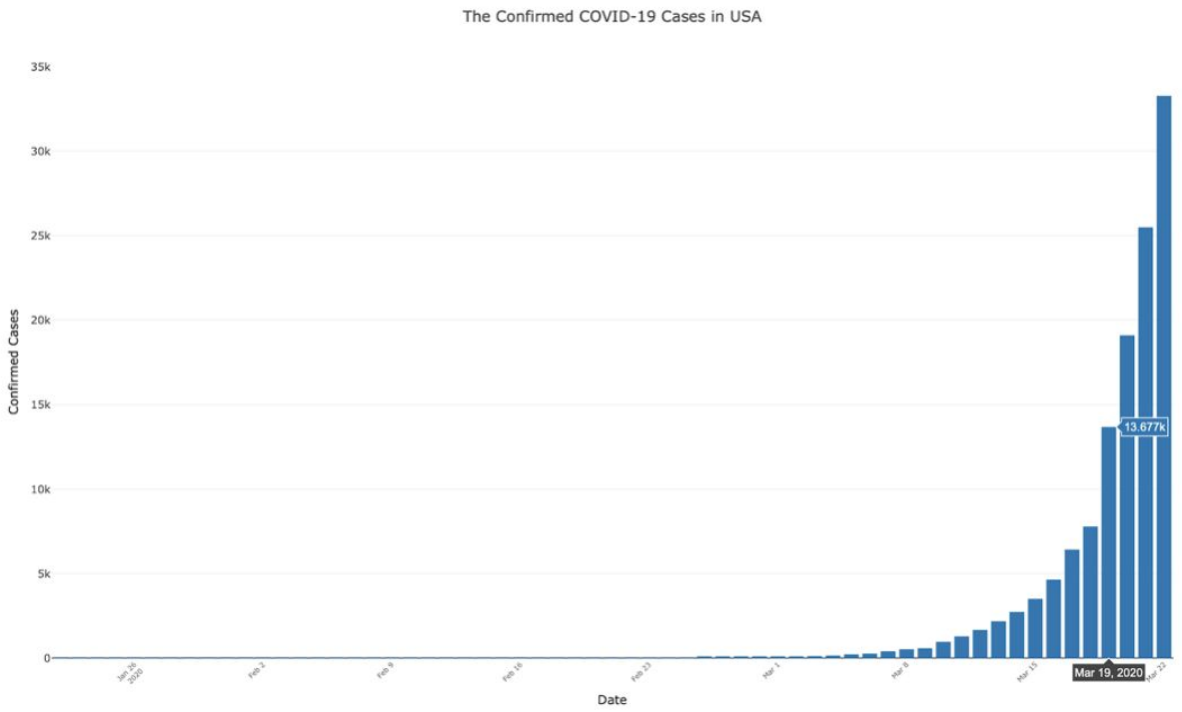
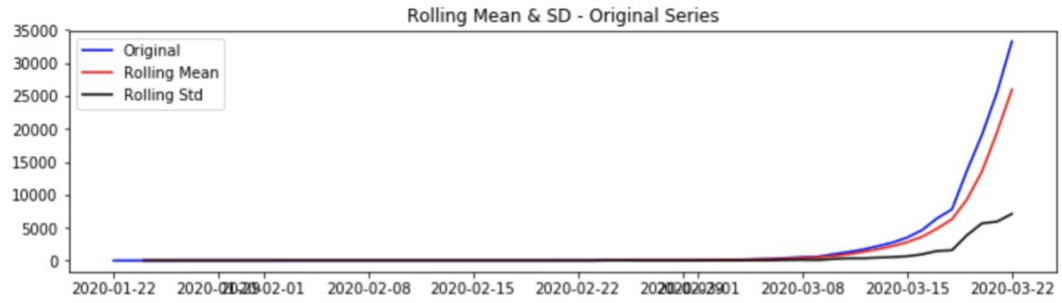


Figure 7: Bar chart of the confirmed COVID-19 cases in the USA from 2020-01-22 to 2020-03-22

3.3.1.2 Checking Stationarity

Before the models are built, the Rolling Statistics and Augmented Dickey-Fuller Test are used to determine whether the time series is stationary. If the rolling statistics exhibit a clear trend (upwards or downwards) and show varying variance (increasing or decreasing amplitude), then the series is very likely not to be stationary [7]. The Augmented Dickey-Fuller test gives a test statistic result about stationarity.

Figure 8 provides the plot for rolling statistics and results of the ADF test on the original data. Both rolling mean (red line) and rolling standard deviation (black line) increase with time. All ADF test statistics are far from critical values, and p-values are greater than all levels threshold. It is easy to conclude that the series of original data is very likely not stationary.




```
 > Is the data stationary ?
ADF Statistics = 3.4268
p-value = 1.0000
Critical Values:
        1%: -3.5714715250448363 - The data is NOT stationary with 99% confidence
        5%: -2.922629480573571 - The data is NOT stationary with 95% confidence
        10%: -2.5993358475635153 - The data is NOT stationary with 90% confidence
```

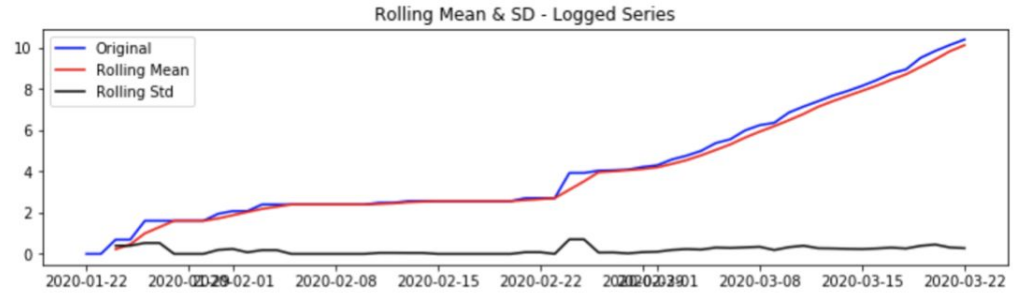
Figure 8: Stationarity test on the original series of USA data. The upper plot is for rolling stationary and lower text messages are the test statistic of the ADF test.

The stationarities on the original series with Rolling Mean, Detrending and Differencing are checked respectively. None of them shows the stationarity.

In this case, the *log* of the series is applied to render stationary. After comparison among all tests, Subtract Rolling Mean on logged data is chosen for modeling.

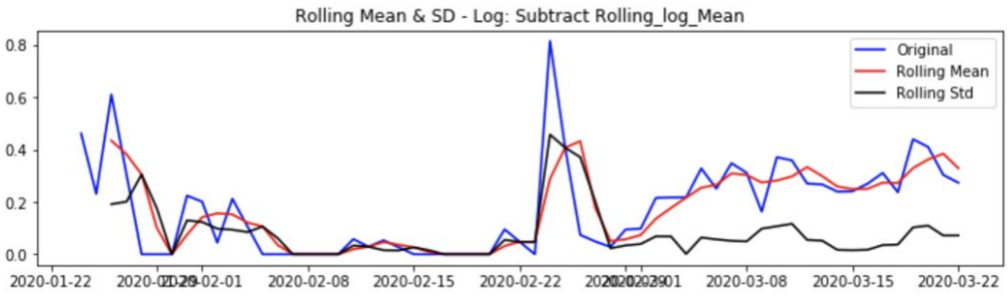
Figure 9 shows the result for logged data. Its Rolling mean (red line) increases with time. All ADF test statistics are far from critical values, and p-values are greater than all levels threshold. The Logged series is very likely not to be stationary. Figure 10 shows

the result for Subtract Rolling Mean on logged data. The rolling mean and standard deviation are approximately parallel and horizontal. The p-value is below the threshold of 0.01, and the ADF statistics are close to the critical values. Therefore, the time series is stationary.




```
 > Is the data stationary ?
ADF Statistics = 2.0852
p-value = 0.9988
Critical Values:
        1%: -3.5443688564814813 - The data is NOT stationary with 99% confidence
        5%: -2.9110731481481484 - The data is NOT stationary with 95% confidence
        10%: -2.5931902777777776 - The data is NOT stationary with 90% confidence
```

Figure 9: Stationarity test on the logged data of USA data. The upper plot is for rolling stationary and lower text messages are the test statistic of the ADF test.




```
 > Is the data stationary ?
ADF Statistics = -4.3615
p-value = 0.0003
Critical Values:
        1%: -3.548493559596539 - The data is  stationary with 99% confidence
        5%: -2.912836594776334 - The data is  stationary with 95% confidence
        10%: -2.594129155766944 - The data is  stationary with 90% confidence
```

Figure 10: Stationarity test on the Subtract Rolling Mean on logged data. The upper plot is for rolling stationary and lower text messages are the test statistic of the ADF test.

3.3.1.3 Parameters Selection

An ARIMA model is noted as ARIMA(p, d, q), where p represents the number of autoregressive parts (AR order), d the order of differencing, and q the number of moving-average terms (MA order).

- Inspection on AutoCorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

By observing the ACF and PACF plots in Figure 11, both of them have lag-1. So 1-order of AR and 1-order of MA are suggested. The difference order might be 1 or 0. The selections for best orders are presented as following procedures.

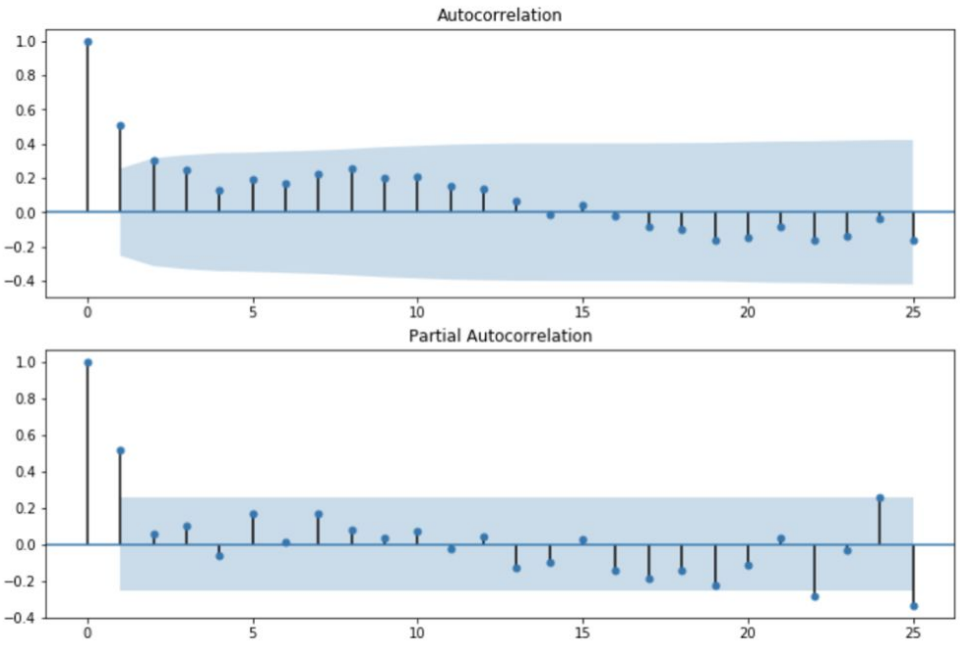


Figure 11: Plots of ACF and PACF for Subtract Rolling Mean on logged data.

- Choosing the differencing order

The models with differencing order of 0 and 1, while keeping AR and MA orders to 0, are fitted. The model fitting summaries and plots are displayed in Figure 12-1. AIC and variance (in yellow panels) are -34.68 and 0.174 for ARIMA(0,0,0), and -37.60 and 0.169 for ARIMA(0,1,0). Model ARIMA(0,1,0) has lower AIC and variance, meaning it is performing better while keeping AR and MA orders equal to 0.

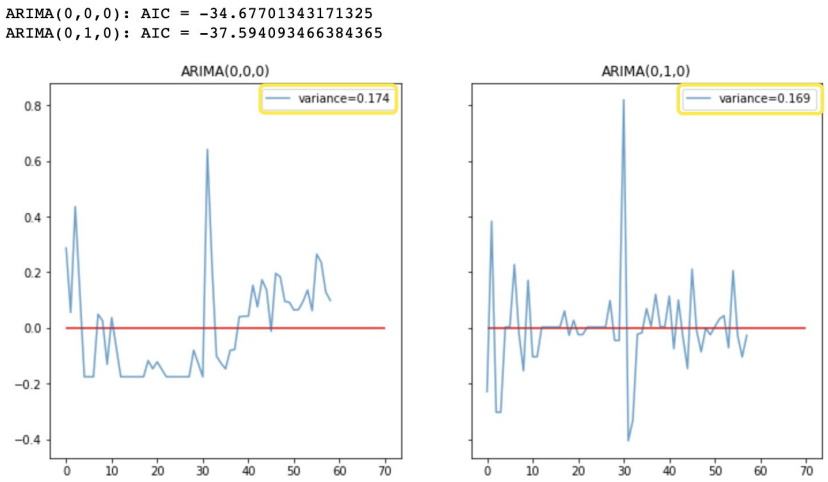


Figure 12-1: Fitting summaries and plots of ARIMA(0,0,0) and ARIMA(0,1,0)

- Choosing the MA and AR order

The same procedures are applied for the MA and AR order selection. Because a 1-order of differencing is already implemented when the logged data is subtracted by rolling mean, which is called Subtract Rolling Mean on logged data, the 0-order of differencing is still used during the MA and AR order selection. After several rounds of selection, the model ARIMA(1,0,0) has the lowest AIC (-51.45) and the second-lowest variance (0.150). Figure 12-2 shows AIC values and plots of fitting models in the last round of comparison.

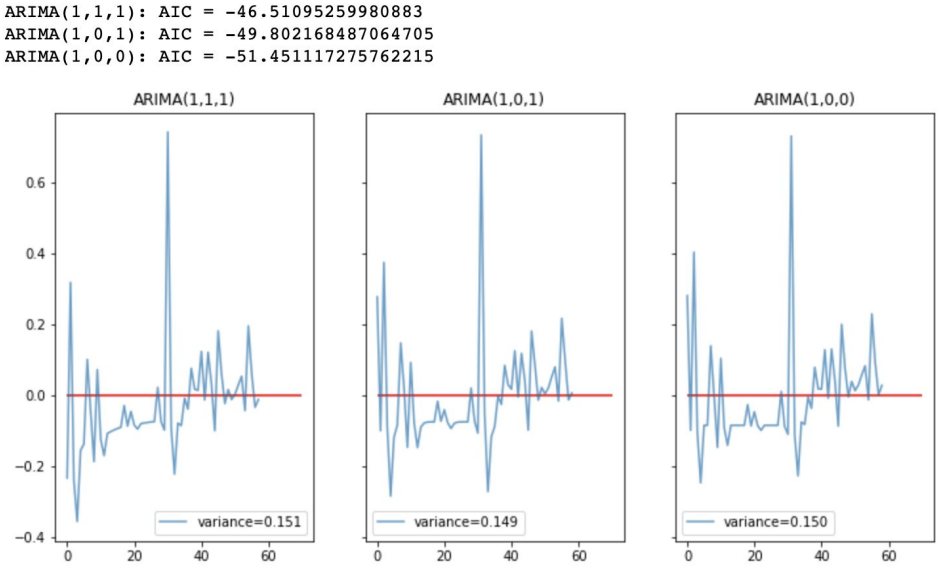


Figure 12-2: AICs and plots of Fitting models

The model ARIMA(1,0,0) of Subtract Rolling Mean on logged data has the best performance among all tested models. It is chosen for the final model to generate a forecast.

3.3.1.4 Model Fitness

- Comparison between the fitted model with Subtract Rolling Mean on logged data

The model ARIMA(1, 0, 0) is fitted. Figure 13 shows the comparison between the fitted values (original line) and Subtract Rolling Mean on logged data (blue line) on a log scale. The RMSE value is 0.1495 at log scale. Table 3 gives the final estimates of parameters. The p-values of constant and AR are close to zero, which indicates all terms in the model are significant.

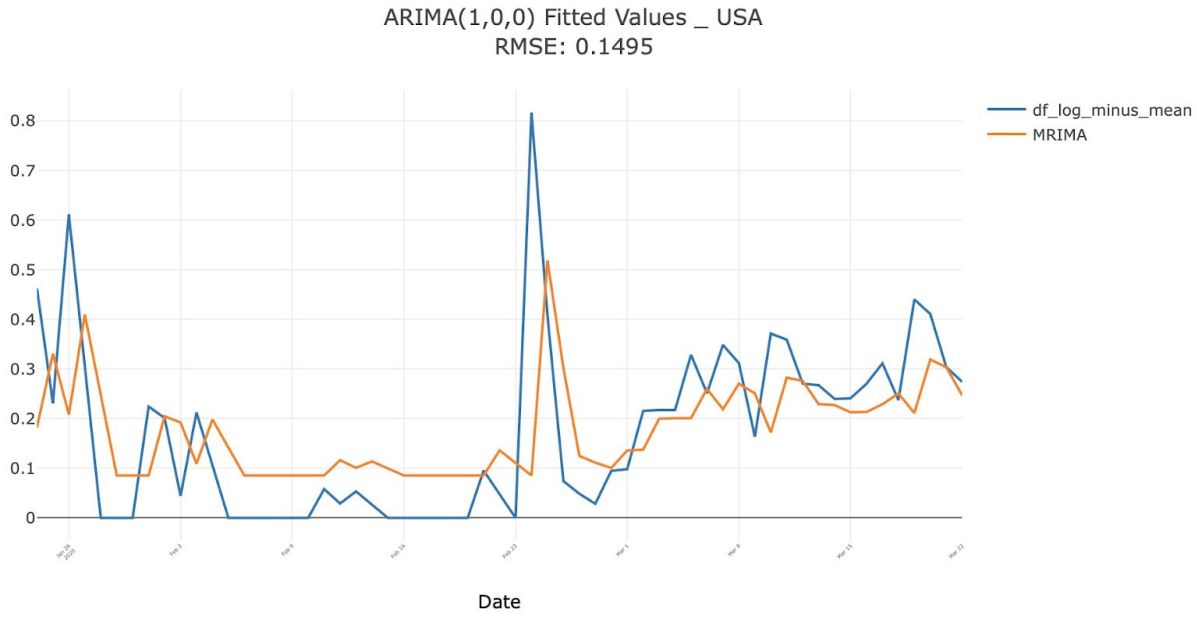


Figure 13: Comparison model fitted values with Subtract Rolling Mean on logged data on a log scale.

Final Estimates of Parameters

```
results_mean.params
```

```
const              0.182107
ar.L1.confirmed    0.530883
dtype: float64
```

```
results_mean.pvalues
```

```
const              0.000034
ar.L1.confirmed    0.000015
dtype: float64
```

Table 3: Final estimates of parameters in the model of ARIMA(1, 0, 0)

- Comparison between confirmed cases and predicted cases

The fitted values from the model ARIMA(1,0,0) are inverted back to the original scale. The procedures include inverting the differencing and applying exponential transform from a log scale. Figure 14 shows the comparison between confirmed and predicted cases on the original scale. The RMSE is 1049 at original scale.

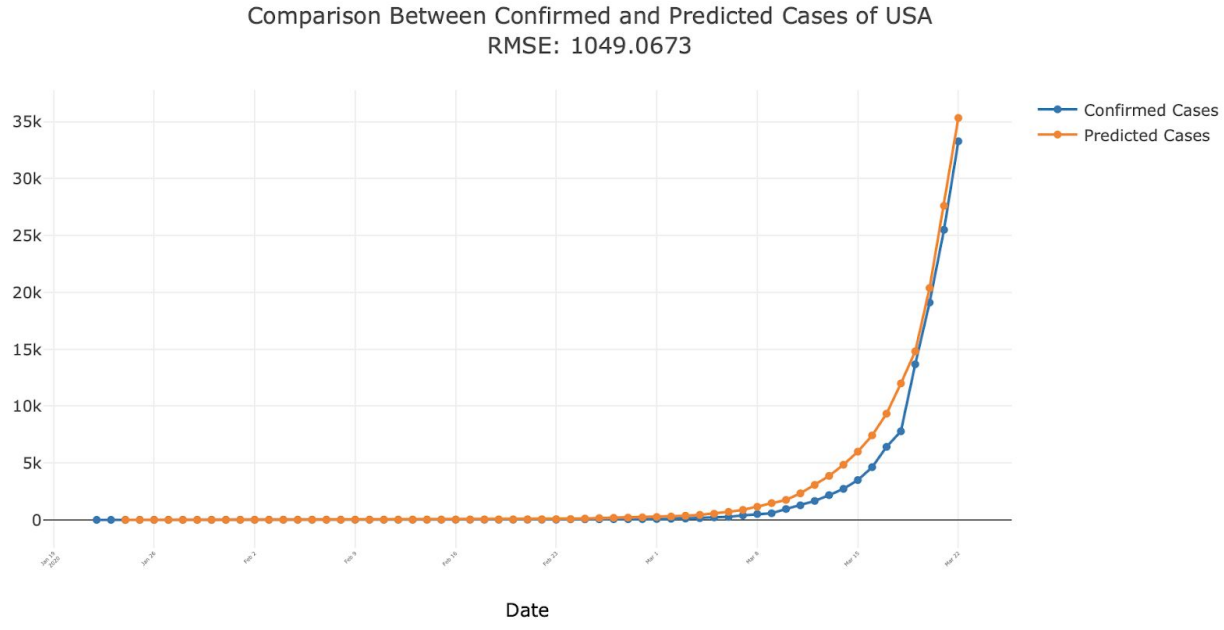


Figure 14: Comparison between confirmed and predicted cases on the original scale.

3.3.1.5 Residual Diagnosis

Residuals are useful in checking whether a model has adequately captured the information in the data.

- Check if residuals are uncorrelated

The ACF and PACF of the residuals are plotted in Figure 15. All the spikes falling into the blue ranges indicates that there is no correlation between residuals and lags of itself. Meanwhile, correlation is also tested using the Durbin-Watson statistic. The Durbin-Watson statistic of the ARIMA model residuals of 2.033 indicates the residuals is no significant correlation in the residuals series, which is consistent with the conclusion from ACF and PACF plots.

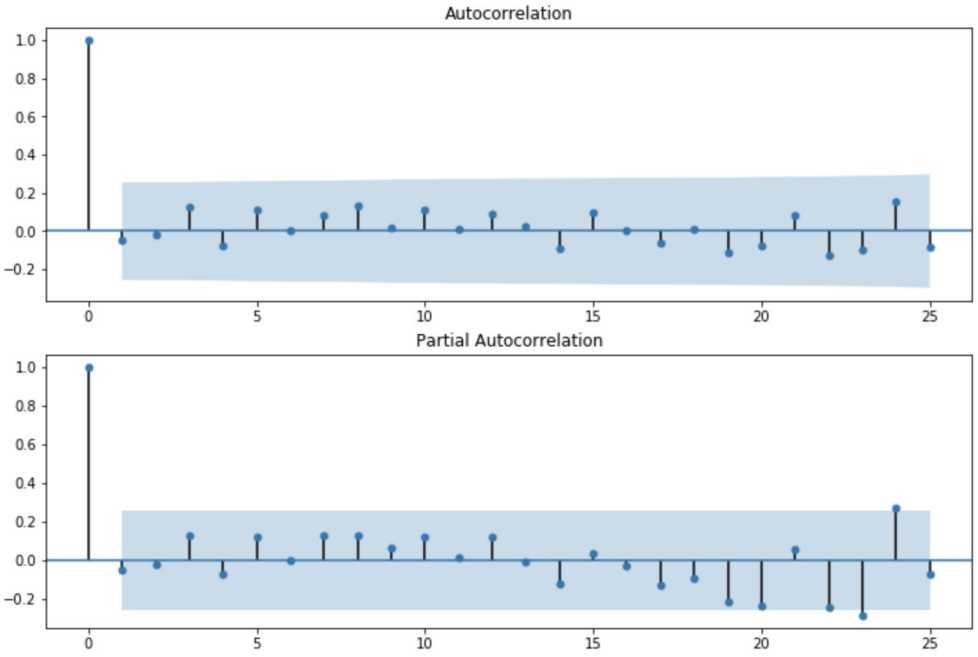


Figure 15: ACF and PACF plots of model residuals

- Check mean and variance

Figure 16 shows the time plot of the residuals. The mean of residuals is close to zero. The variation of the residuals stays much the same across the historical data, except one outlier. Therefore the variance can be treated as constant.

```
mean of the residuals: -0.0025
```

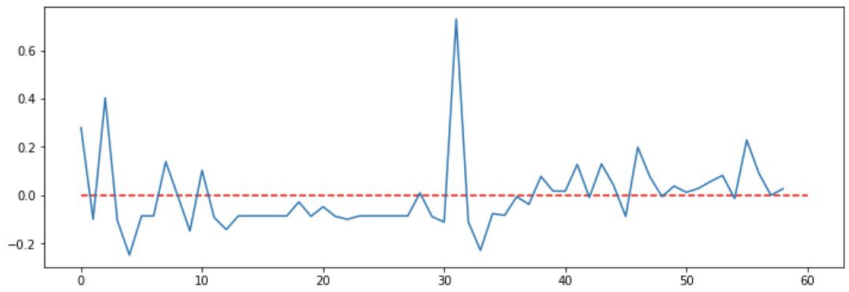


Figure 16: Time plot of the residuals

- Check normality of residuals

Histogram (left) and qqplot (right) in Figure 17 shows that the normality of residual is not perfectly matched. The right tail seems a little too long, and there is a little aperture in the middle part of the qqplot. The effect of outliers might be the reason that causes the non-normality of residuals.

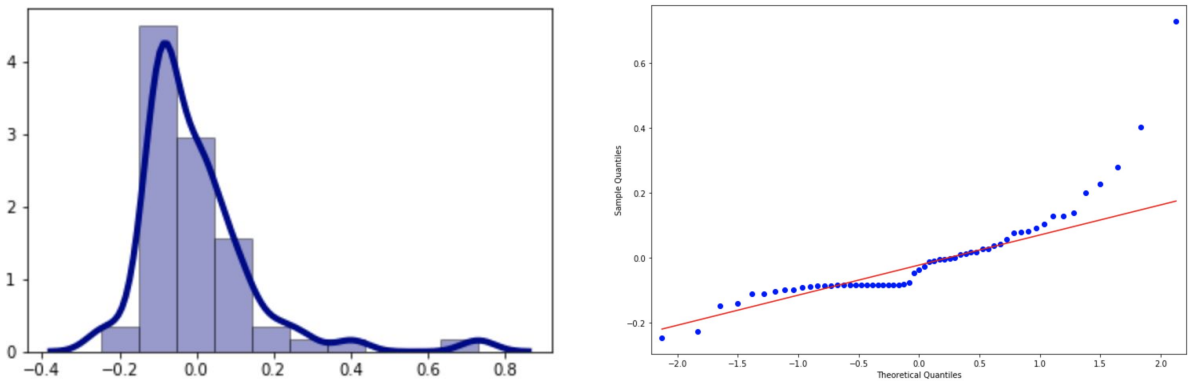


Figure 17: Histogram (left) and qqplot (right) of the residuals for normality check.

The assumption of prediction intervals is a normal distribution. As a consequence of the non-normality of residuals, forecasts from the model are probably quite good, but the prediction intervals may be inaccurate.

3.3.1.6 Forecasts

- In-Sample forecast

Here 12 forecasting steps in-sample are chosen. The forecast is made from March 10-March 22. The algorithm is forced to use its forecasted values as the lagged values for forecasts inside the sample series by setting the '*dynamic*' keyword as True. Figure 18 shows the in-sample forecasts. The trend of forecast (orange line) follows the trend of actual series (blue line) during the forecast period. And all the actual values fall in the confidence bands which are shaded as gray.

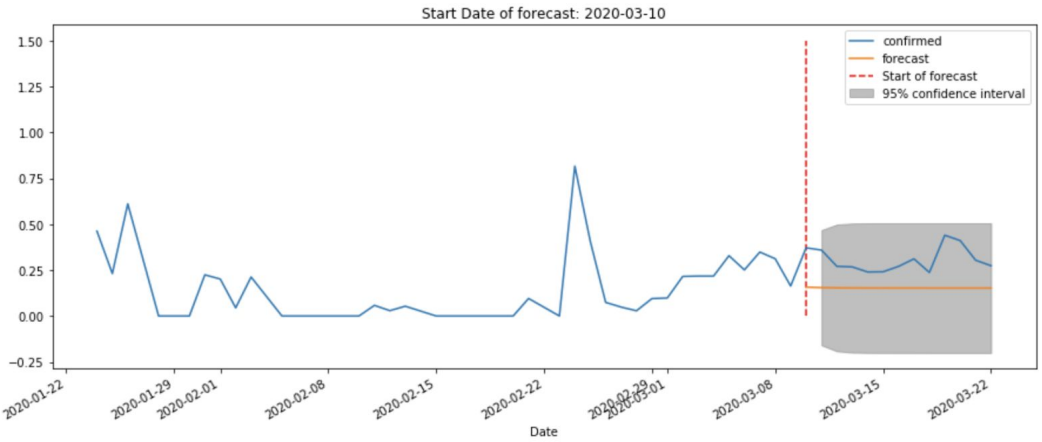


Figure 18: In-Sample forecast from 2020-03-10 to 2020-03-22.

- Out-of-sample Forecasts on a log scale

A 9-day forecast is produced. The forecast is made from March 23 - March 31, the last day of March. Figure 19 shows the forecast curve and its confidence bands. The trend during this 9-day is also keeping the same rate with a little bit of decreasing, which indicates that the exponential-likely increment during this period is expected.

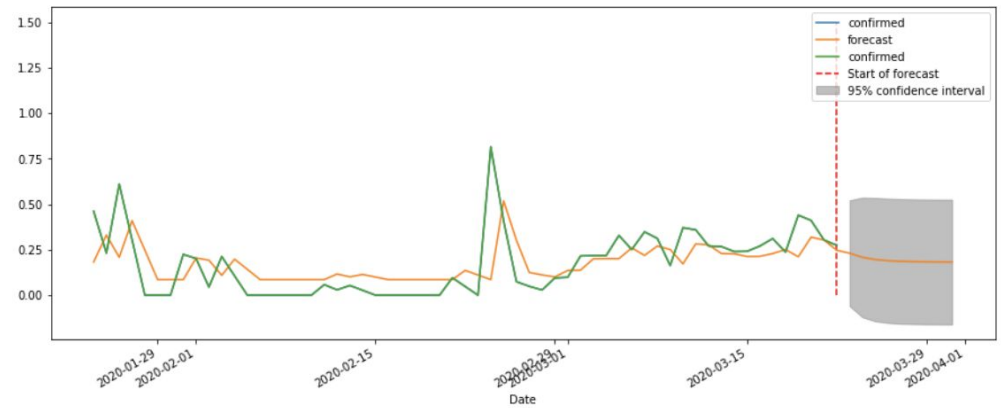


Figure 19: 9-day Forecast from 2020-03-22 to 2020-03-31 on a log scale.

- Convert forecast from log scale back to the original scale

Similar to the previous procedure done for model fitting, the forecasted values are converted back to the original scale. The procedures include inverting from the differencing and applying exponential transform from a log scale. Figure 20 shows a 9-day forecast of confirmed cases in the USA from March 23 to March 31, the end of the month. Table 4 lists the forecasted values of these 9 days. The trend curve is so sharp that there is no sign the increasing rate is going to slow down and there is even no hope to see a turning point in the future 9 days. Each day more and more new confirmed cases are expected. The predicted confirmed cases are soaring to up to 200K at the end of March. It is nearly 6 times of 33k which is the confirmed case on March 22, last day in the dataset. This rings a big alarm that the US government has to take the situation more seriously and must take quick and effective measures to prevent the COVID-19 from spreading.

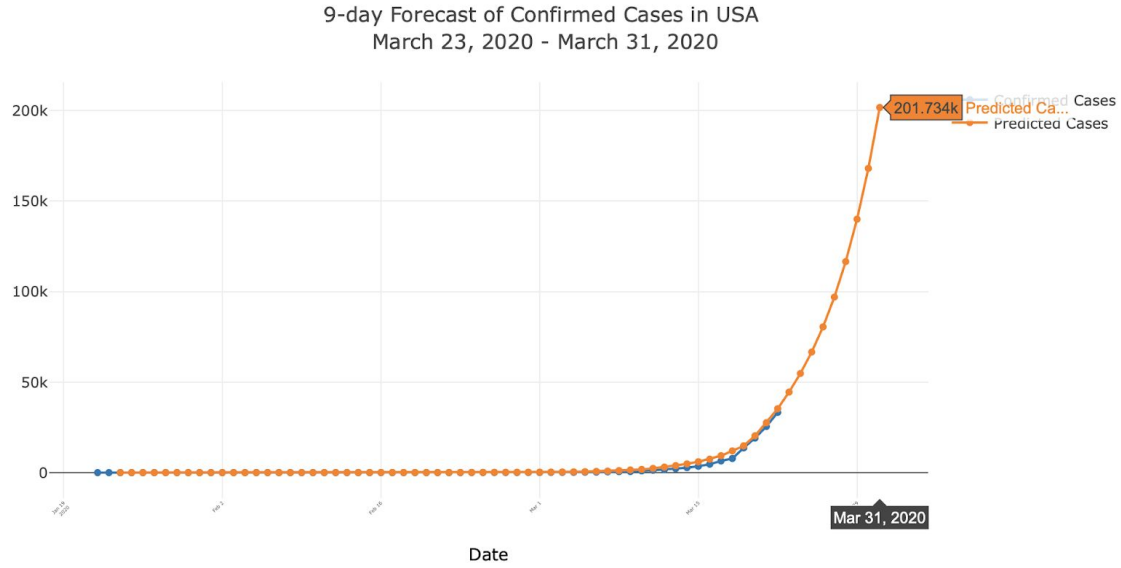


Figure 20: Plot of 9-day Forecast of confirmed cases in the USA from March 23 to March 31

	Predicted
2020-03-23	44493
2020-03-24	54778
2020-03-25	66628
2020-03-26	80521
2020-03-27	96979
2020-03-28	116589
2020-03-29	140029
2020-03-30	168096
2020-03-31	201734

Table 4: 9-day Forecast Values in the USA

3.3.2 ARIMA model for Italy

3.3.1.1 Data visualization of the original Italy data

Italy is the hardest-hit country in Europe. it has the second most confirmed cases by March 22. Figure 21 provides a bar chart of confirmed coronavirus cases in Italy. It shows that the number of cases is as high as 59k on March 22, which is 33k in the USA. However, by comparing both plots, we can find that the USA has a higher increasing rate than Italy, which implies that the cases in the USA will exceed Italy very soon.

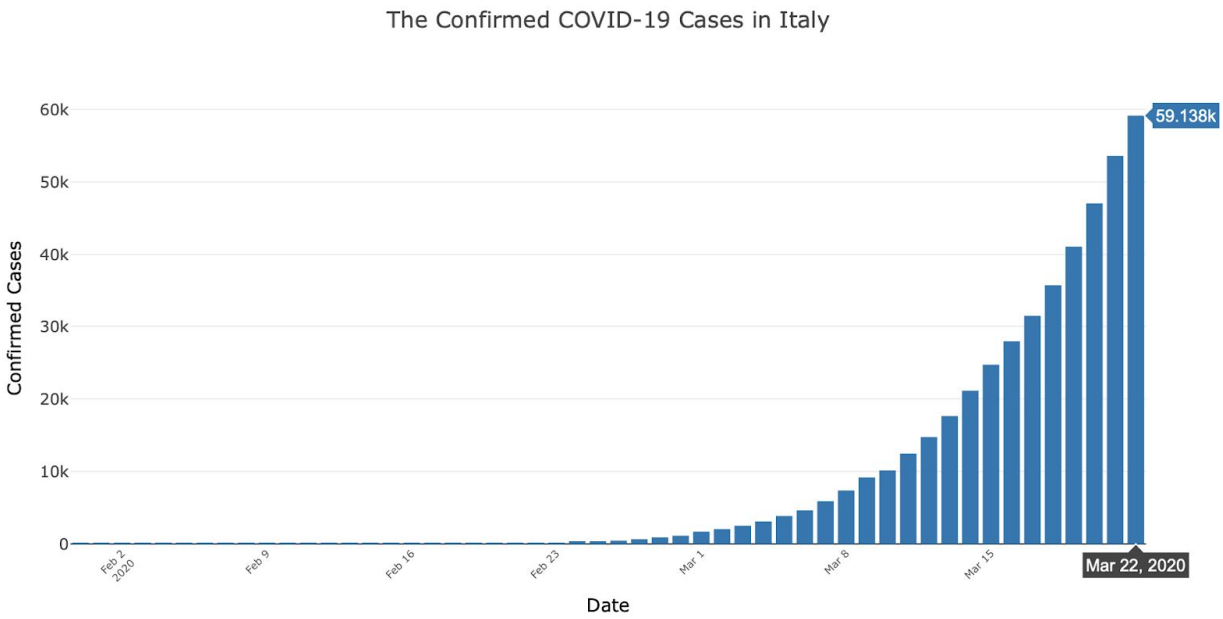


Figure 21: Bar chart of the confirmed COVID-19 cases in Italy from January 22 to March 22

3.3.1.2 ARIMA model selection

The procedures for the ARIMA model selection are the same as mentioned for the USA ARIMA model. The model for Italy is simpler since the series does not need to do *log* transformation. The ARIMA(3, 1, 0) on the original series is chosen for the prediction.

3.3.1.3 Model Fitting

Figure 22 is the graph of the comparison between actual values and fitted values. The model RMSE is 2102.0. Figure 23 plots the comparison between confirmed and predicted values after inverting 1-order differencing back to the original scale. Its RMSE is 1131, which is very close to the RMSE of 1049 for the USA ARIMA model. Both models have similar prediction accuracy.

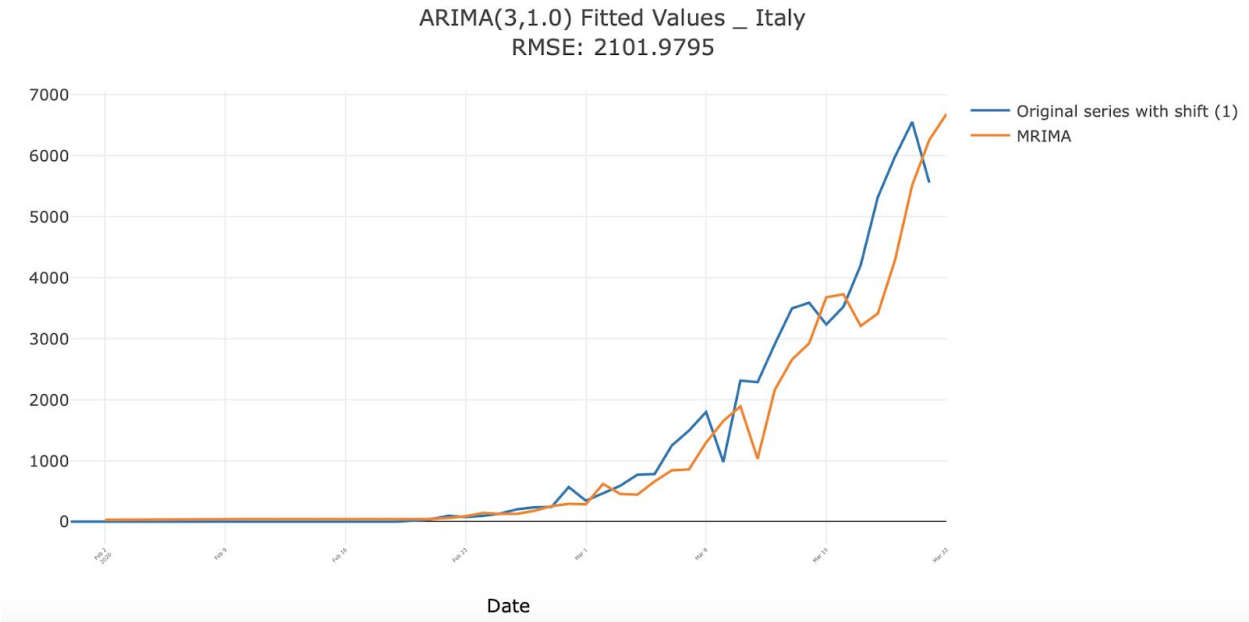


Figure 22: Comparison between actual values and fitted values.

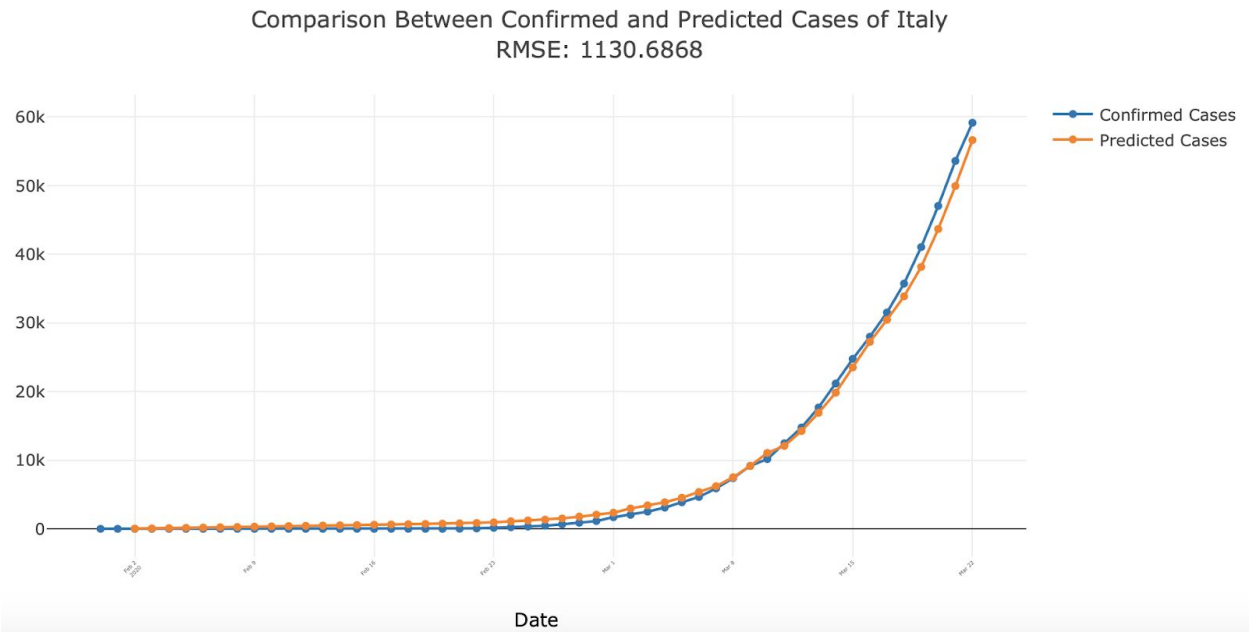


Figure 23: Comparison between confirmed and predicted cases

3.3.1.4 Forecast

Figure 24 shows a 9-day forecast of confirmed cases in Italy from March 23 to March 31, the end of the month. Table 5 lists the forecasted values of these 9 days. The good sign is that the increasing rate is about to slow down, even though it's not very much. The new cases each day are expected to get less and less in the future for 9 days. The turning point will appear if the trend can be kept.

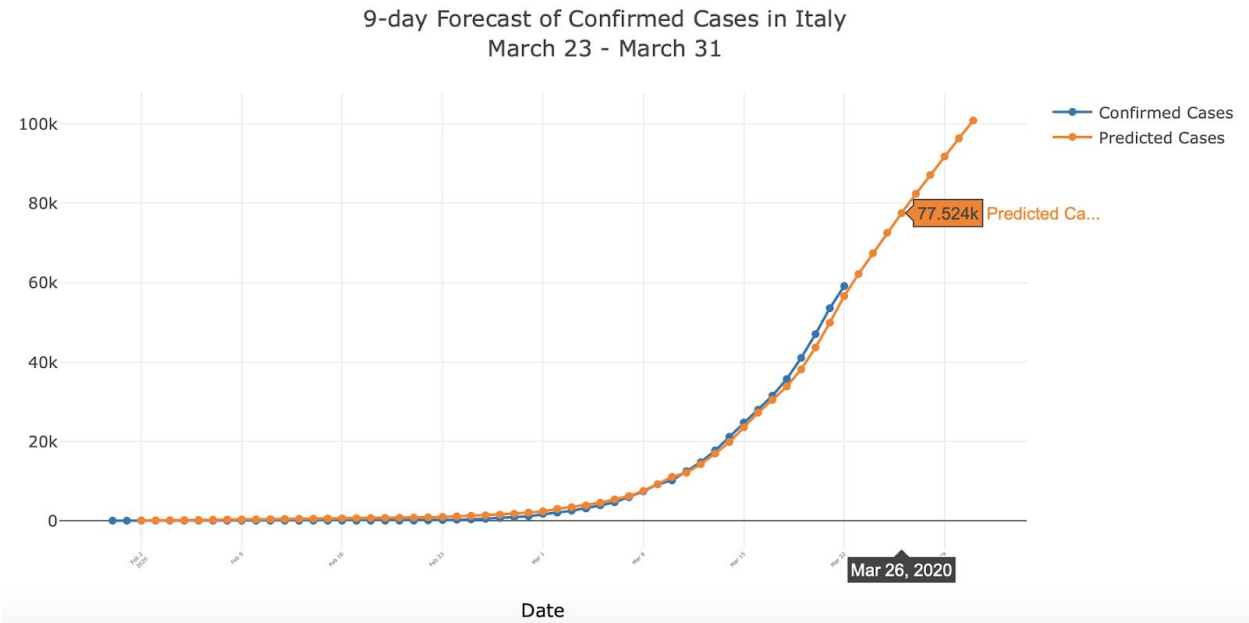


Figure 24: Plot of 9-day Forecast of confirmed cases in Italy from March 23 to March 31

	predicted
2020-03-23	62188
2020-03-24	67414
2020-03-25	72555
2020-03-26	77524
2020-03-27	82396
2020-03-28	87152
2020-03-29	91815
2020-03-30	96384
2020-03-31	100866

Table 5: 9-day Forecast Values in Italy

Looking back to the forecasts for the USA, just as we anticipated, the expected cases in Italy will be exceeded by the USA 4 days later on March 26. At the end of March, the expected confirmed cases in Italy is 101k, half of the cases in the USA, which is 201K.

4. Discussion

4.1 Technical Point of View

On the technique side, the online Kalman filter algorithm provides very powerful and accurate 1-day forecasts. This algorithm builds one model which works on all the countries/states simultaneously with high accuracy. It is very efficient and requires low maintenance. Since it is an online algorithm and each prediction summarizes the previous online predictions, there is no overfitting or bias [6].

It is an exciting experience to run an online Kalman algorithm. It runs in real-time. All the operations, such as file import, data preprocessing/processing, plots of data visualization, etc, can be updated promptly when the program is re-run. One drawback encountered in this work is that the data processing procedure in the Python script has to be revised if there are changes on the format of the data source.

Analyzing series is a fascinating job for predicting the future and dealing with uncertainty, and ARIMA is a very powerful model for forecasting time series data. In this work, the ARIMA algorithm is applied to build models in the USA and Italy. Furthermore, the models are used to generate short term forecasts. The fitness of both models is very good. Their RMSEs are as low as 1100.

As the confirmed cases on March 23 are available right now, the predicted values from Kalman filter and ARIMA are listed in Table 5 for comparison. For the USA cases forecasting, the prediction value from the USA ARIMA model is very close to the true value, while the prediction error of the Kalman model is relatively high. For Italy cases, the Kalman model gives a better forecast than the Italy ARIMA model.

The Confirmed Cases and Predicted Cases on MARCH 23				
USA		Confirmed	Kalman Predicted	ARIMA Predicted
		43781	40674	44493
	Prediction Error		3107	-712
Italy		Confirmed	Kalman Predicted	ARIMA Predicted
		63927	65076	62188
	Prediction Error		-1149	1739

Table 5: The Confirmed Cases and Predicted Values on MARCH 23

4.2 Insights into Current Events

- The Chinese lockdown measure and South Korea's expansive diagnostic capacity at scale and extensive contact tracing are the successful strategies that can be used as references when each country handles its emergency status.
- Instead of China, Europe is the epicenter of the pandemic in the middle of March.
- Like Hubei province, New York becomes the epidemic center of the USA. The very sharp trend of New York implies its future is going to be very tough.
- Italy and the USA are the top 2 and 3 with the highest confirmed cases on March 22. However, both of them are expected to surpass China around March 26, if the trend for China keeps steady. At the same time, the confirmed cases in the US are very likely to exceed both Italy and China, unfortunately becoming number 1.
- At the end of March, the confirmed cases in Italy will reach 100k, and in the USA it is about to soar to up to 200k, double the number of its instant follower Italy.
- Hope emerges from the 9-day forecast for Italy. Its increasing rate is slowing down, even though it is not substantial. The turning point is going to arrive, hopefully soon, unless an unforeseeable event occurs.
- The 9-day forecast for the USA rings an emergency siren. The trend is so sharp that there is no sign the increasing rate is expected to slow down. It is unsure when the turning point will arrive.

5. Future Work

It is worth taking time to implement *Auto ARIMA* in future research. Although ARIMA is a very powerful model, the data preparation and parameter tuning processes end up being time-consuming [10]. Auto ARIMA makes this task simple as it eliminates the steps for making series stationary and determining the values of p and q. The further goal is to make an online Auto ARIMA algorithm. After the Auto ARIMA model works functionally, the online train should be feasible.

Appendix A

Kalman Filter

The Kalman filters are used to estimate states based on linear dynamical systems in state space format [8].

The process model of state from time k-1 to time k as

$$x_k = Fx_{k-1} + Bu_{k-1} + w_{k-1}$$

Where F is the state transition matrix

B is the control-input matrix

w_{k-1} is the process noise vector and $w_{k-1} \sim N(0, Q)$

The measurement model at current time step k as

$$z_k = Hx_k + v_k$$

Where z_k is the measurement vector

H is the measurement matrix

v_k is the measurement noise vector and $v_k \sim N(0, R)$

The information of the system in Kalman filter is described by F, B, H, Q *and* R.
The Kalman filter algorithm is summarized as follows [8]:

Prediction:

Predicted state estimate	$\hat{x}_k^- = F\hat{x}_{k-1}^+ + Bu_k$
Predicted error covariance	$P_k^- = FP_{k-1}^+F^+ + Q$

Update:

Measurement residual	$\hat{y}_k = z_k - H\hat{x}_k^-$
Kalman gain	$K_k = P_k^- H^T(R + HP_k^-H^T)^{-1}$
Updated state estimate	$\hat{x}_k^+ = \hat{x}_k^- + k_k\hat{y}_k$
Updated error covariance	$P_k^+ = (I - k_k H)P_k^-$

In the above equations, the hat operator means an estimate of a variable. - and + denote predicted and updated estimates, respectively.

In the work, the one-dimensional predicted value (p) and its change rate (v) comprise the state vector:

$$x = [p, v]^T$$

The state in time k as:

$$x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} x_{k-1} + \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} \tilde{u}_{k-1}$$

Where $u_{k-1} = \tilde{u}_{k-1} + e_{k-1}$, and $e_{k-1} \sim N(0, \sigma_e^2)$

And the covariance matrix of the process noise as:

$$Q = \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} \sigma_e^2 \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix}^T = \begin{bmatrix} \frac{1}{4}(\Delta t)^4 & 0 \\ 0 & (\Delta t)^2 \end{bmatrix} \sigma_e^2$$

So the process model as:

$$x_k = Fx_{k-1} + Bu_{k-1} + w_{k-1}$$

Where

$$F = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \quad and \quad B = \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix}$$

$$w_{k-1} \sim N(0, Q)$$

The measurement model with measurement noise v_k as:

$$z_k = Hx_k + v_k$$

Where $H = I_{2*2}$ and $v_k \sim N(0, R)$

References

[1] BBC News March 13, 2020 https://www.bbc.com/news/world-us-canada-51882381

[2] Kimberlyn Roosa, Yiseul Lee, Ruiyan Luo, Alexander Kirpich, etc *Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13-23, 2020* Journal Clinical Medicine, 2020-9

[3] Kalman Filter
https://en.wikipedia.org/wiki/Kalman_filter

[4]Time Series in Python — Exponential Smoothing and ARIMA processes
https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788

[5]How to Create an ARIMA Model for Time Series Forecasting in Python
https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
[6]Using Kalman Filter to Predict Coronavirus Spread, Ran Kremer
https://towardsdatascience.com/using-kalman-filter-to-predict-corona-virus-spread-72d91b74cc8

[7]Time Series in Python — Exponential Smoothing and ARIMA processes
https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788

[8]Introduction to Kalman Filter and Its Applications By Youngjoo Kim and Hyochoong Bang, November 5th, 2020
https://www.intechopen.com/books/introduction-and-implementations-of-the-kalman-filter/introduction-to-kalman-filter-and-its-applications

[9] Forecasting Principles and Practice https://otexts.com/fpp2/residuals.html

[10]Build High-Performance Time Series Models using Auto ARIMA in Python and R
https://docs.google.com/document/d/17hFKfXoaOOSV-Q55sLKTSVzzD1JqEhLb1raqekehLjQ/edit#