Forecasting Covid-19: Multi-Variate Time Series

Ong Zheng Kai

P2004048

DAAA/2A04

**Abstract**

It has been almost 2 years since the COVID-19 pandemic begun. Yet, everyone around the world is still battling the virus hard. To date, there are multiple measures and tools in place for government agencies to use to bring the virus under control. One of these tools is applying stricter and tighter restrictions in the country. Applying these tight measures in a timely fashion is also vital before the virus spirals out of control. The question to always ask is, when should these tighter measures take effect. To know the when, is always a good idea to forecast the potential number of cases and deaths. This forecast could also be made public as a cautionary note and be aware when the virus is about to spike or for business owners to better prepare before tighter measures are put in place. Considering we are forecasting COVID-19 infections and deaths, we will be applying multivariate time series forecasting with VARMA. The root mean squared error is compared to 2 ARIMA models. With an extra tool, forecasting COVID-19, fighting the virus might just be easier.

1. **Introduction**

In this day and age, Artificial Intelligence (AI) is used to solve most problems we face that otherwise cannot be solved by humans. For example, an advanced technique to detect COVID-19 was made possible with a model that can detect asymptomatic COVID-19 infections with recordings of one’s cough (Jennifer, 2020). Also recently, there has also been a breakthrough drug found by a group of Singaporean researchers with the help of AI that is able to treat patients infected with COVID-19 6.5 times more effective than the Ebola drug (Audrey, 2020). These articles highlight the vital role AI plays in the medical industry.

In general, AI has helped to better detect the virus, find a more effective treatment to treat the virus. How about forecasting the trajectory of the new cases and new deaths? In April 2020, Singapore saw an influx of cases in our dorms. On April 5th, the number of new cases exceeded 100. From then on, cases continue to spiral out of control, as new cases exceed to 200, then 300. By April 16th, a record 728 new cases were reported. Within a span of 11 days, the number of new cases had increased by a factor of 7 times. The circuit breaker measure commenced on April 7th (Michael, 2020). However, by then, new cases were already in their hundreds. This setback taught us the importance of timely implementation of strict measures. However, one could only imagine how the number of new cases would have spiked.

Learning from the lesson of April 2020, we can also see the importance of being ahead of the virus. With that, we must forecast the virus’s trajectory, so that measures could not only be imposed earlier, but for shorter periods of time. This paper is going to explore multi-variate time series with the use of the Vector Auto Regressor Moving Average (VARMA) model to forecast the new COVID-19 infections and deaths. We will also be comparing the VARMA model to two ARIMA models, one for predicting new COVID-19 infections, whilst the other will predict the number of new deaths in Singapore.

1. **Related Works**

With COVID-19 still on the rise, some researchers have jumped on the band wagon are forecasting COVID-19. As of now, there are dozens of models forecasting COVID-19. However, these projects are based in the United States. These projects make 4 weeks projections, however, only a handful had accuracy above 70% compared to the baseline model of Texas Tech University. One of the most notable projects was covid19-projections.com. It was the best performing model by far as used a Susceptible, Exposed, Infectious, Recovered (SEIR) model to simulate and forecast the infections (covid19-projections.com, n.d.). Sadly, the project is no longer maintained.

The CDC has even begun to show COVID-19 forecast for the United States. It uses multiple COVID-19 prediction models and gives a 4-week projection. Another researcher made a model for India. He experimented with various models such as ARIMA and Extra Tree Classifier. It’s ARIMA model was able to achieve RMSE of 253.6 (Deepak, 2021).

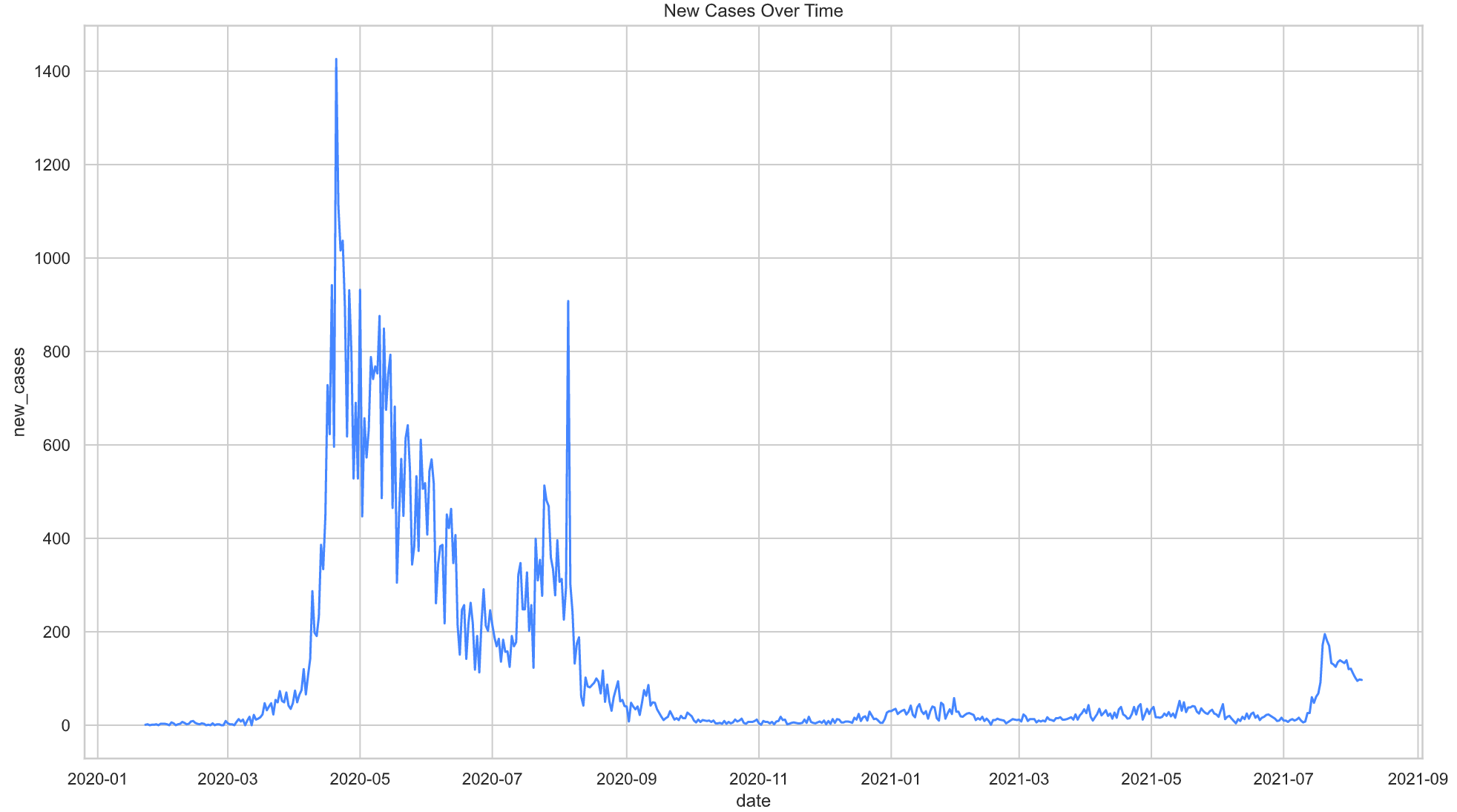
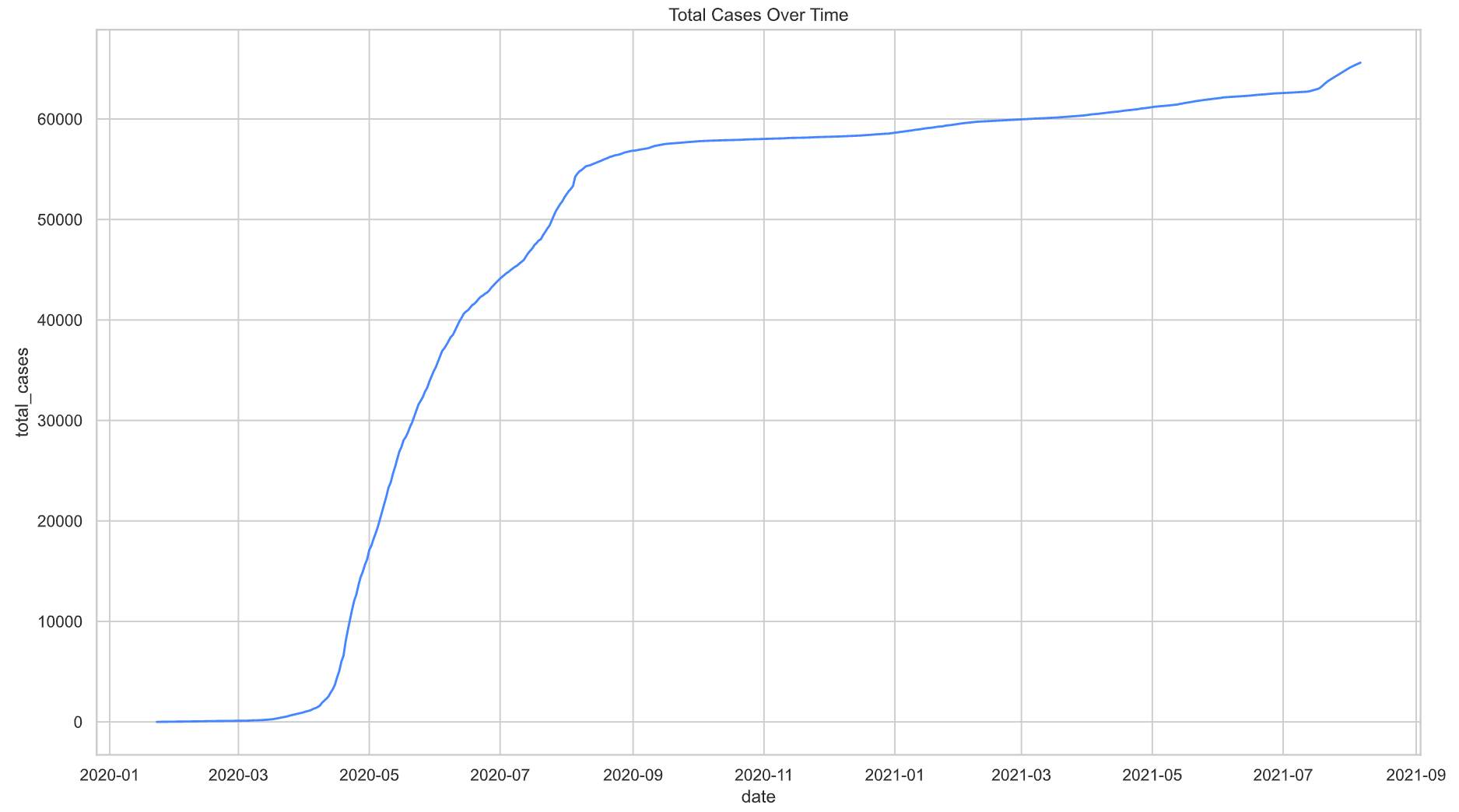
Locally, there are no known AI model that is able to forecast and predict COVID-19 cases, let alone COVID-19 deaths.

1. **Dataset**
   1. **Understanding the data**

The Dataset comes from Our World in Data’s GitHub repository. The data is update daily and we will be reading the data from the CSV download link. The dataset contains COVID-19 related data for 194 countries. However, some data are not available in Singapore, like number of hospital patients, number of ICU patients and COVID-19 testing. Most of which does not affect daily cases, other than the COVID-19 testing. For testing our model, we will take the latest 4 weeks of data.

As of 8th August 2021, there are a total of 561 rows of data. This means, we will have 533 rows of data for training, and 28 rows of data for testing. Our variables will be “new\_cases” and “new\_deaths”. To improve our prediction, we will be including the following features: “people\_vaccinated”, “people\_fully\_vaccinated”, “stringency\_index”. The data is relatively clean. However, cleaning has to be done for the features. This is because, these features are periodically updated. The dates without any updates will be left blank.

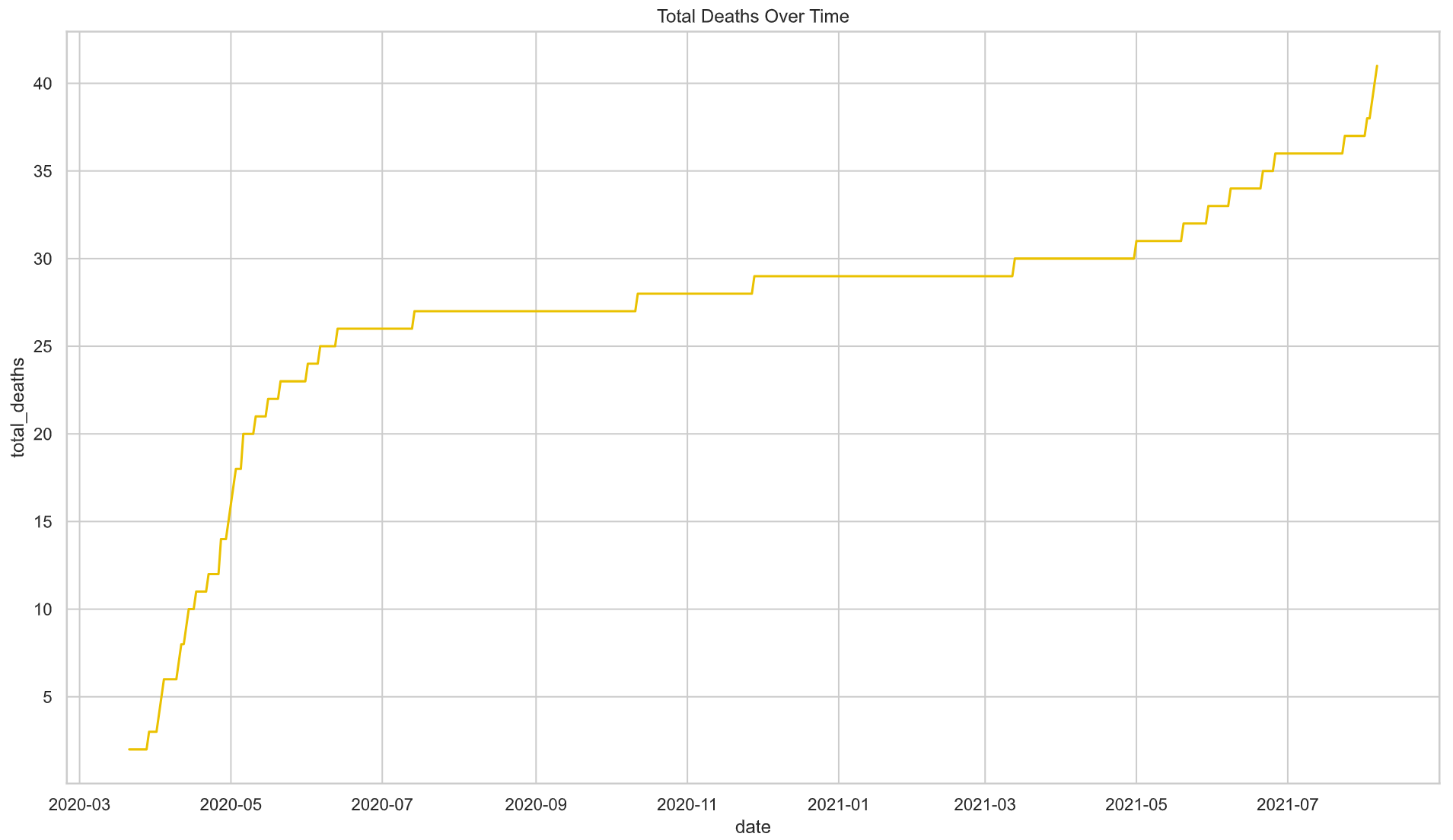
* 1. **Exploratory Data Analysis**

****

|  |  |
| --- | --- |
| Fig 1. Total Cases Over Time | Fig 2. New Cases Over Time |

In Figure 1, we can see cases climbing steeply from April 2020 to August 2020. The cases then climb at a less steep gradient and during late July in 2021, we once again see a slight peak of cases.

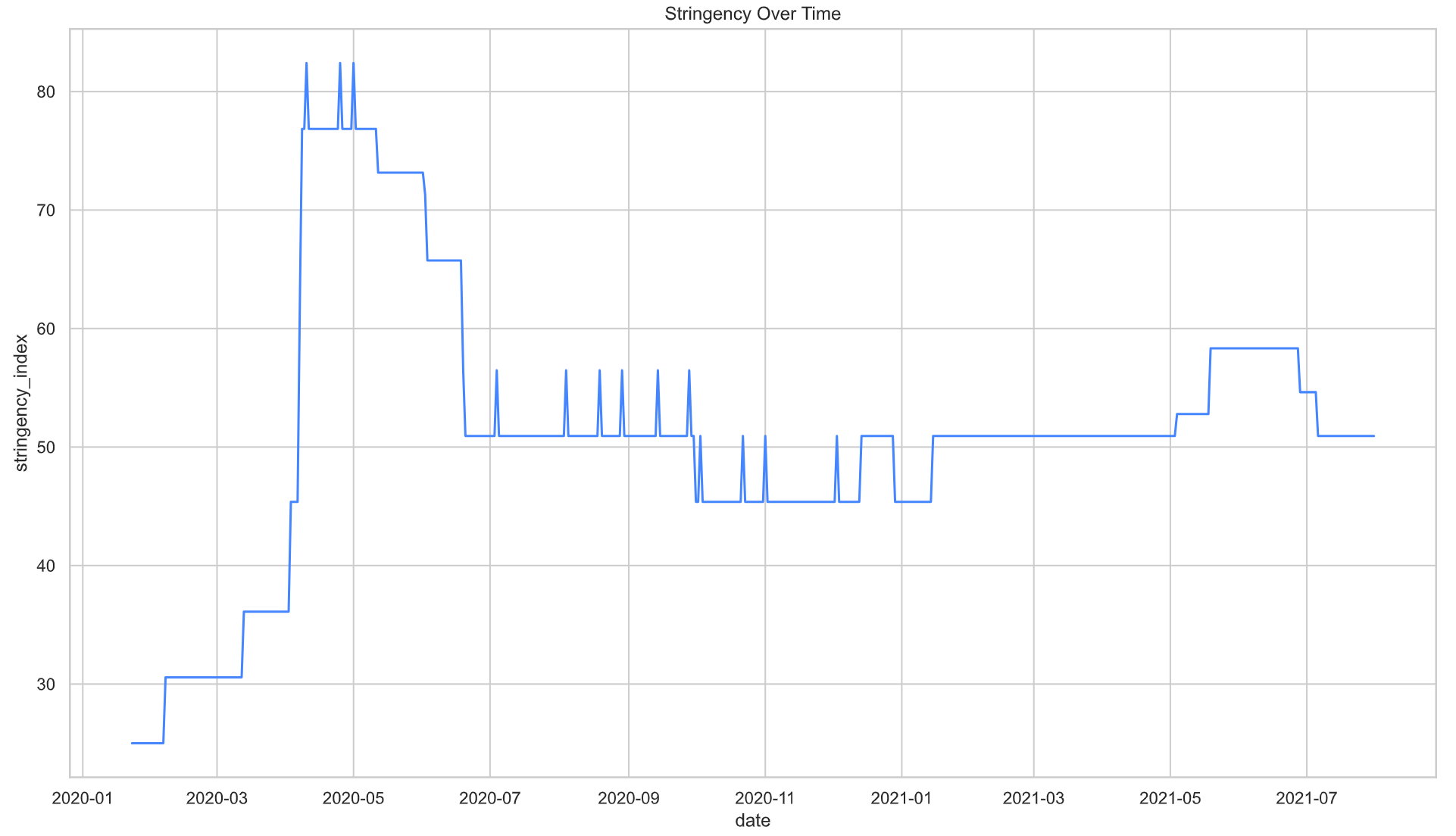
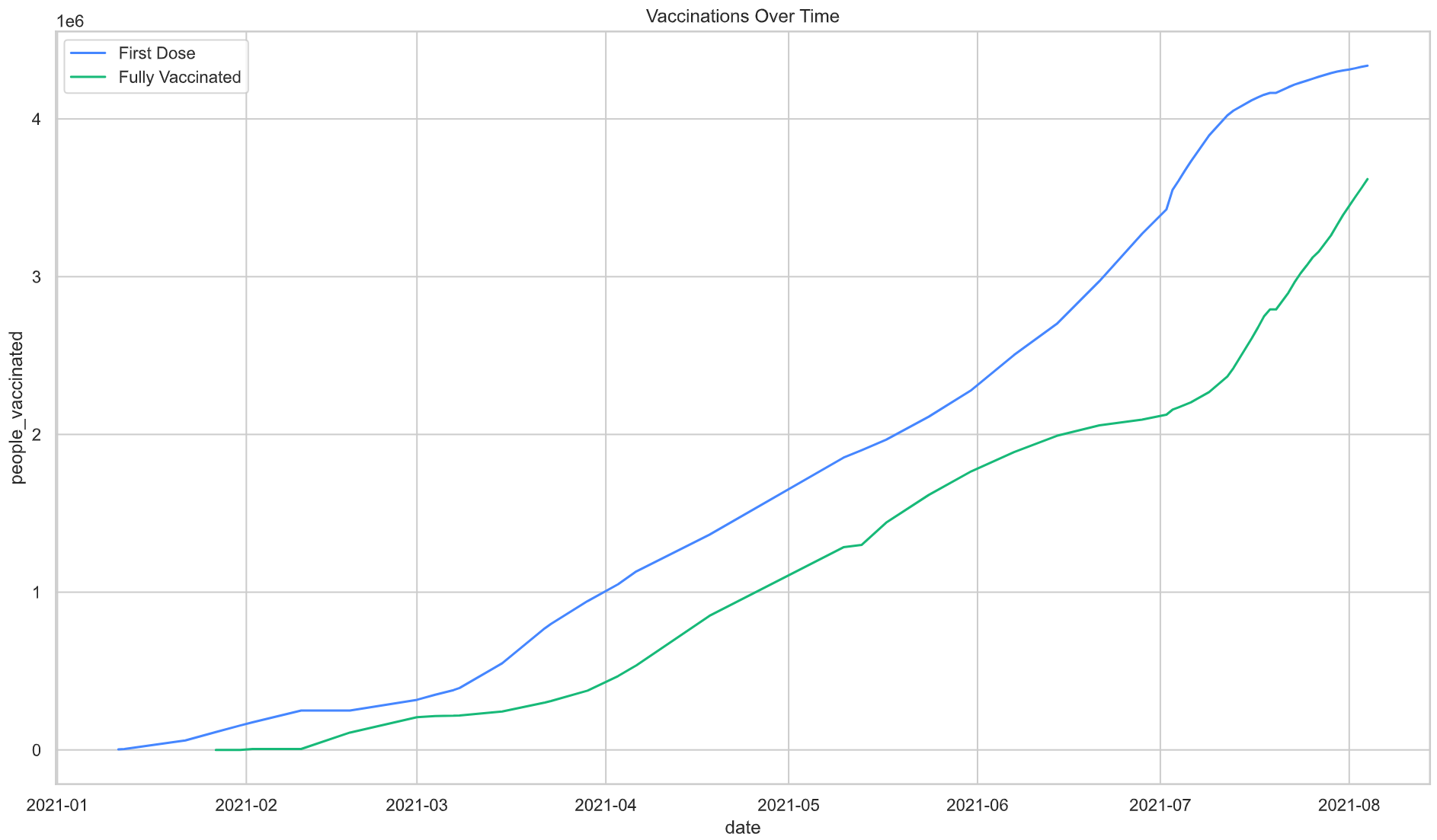
This analysis corresponds to Figure 2. We can see that from April 2020 to August 2020, the number of new cases is above 200, peaking at over 1400. During late July 2021, we can see a sudden rise in new cases and peaking at around 200 cases before dropping.



|  |
| --- |
| Fig 3. Total Deaths Over Time |

In Figure 3, we can see that the number of deaths over time is correlated with the total number of cases over time. In months April 2020 to August 2020, the number of deaths increased steeply. This sudden rise of deaths is correlated with the total number of cases observed in Figure 1.

From Figure 1 and 3, we can conclude that the total number of cases is correlated with the total deaths over time.

****

|  |  |
| --- | --- |
| Fig 4. Stringency Over Time | Fig 5. Vaccinations Over Tim |

Figure 4 depicts the tightness of measures implemented by the Singapore government over time. It is observed that the stringency does indeed correspond to the level of severity of the virus as seen in Figure 2. The stringency index also seems to match the safety measures imposed by the Government. For example, in April 2020, the government imposed the Circuit Breaker. This was by far the strictest measure the government have imposed as in corresponds on the Stringency index.

Singapore’s current population is 5.69 million (singstat.gov.sg, 2021). Based on Figure 5, Singapore has fully vaccinated about 3.6 million of the population, which is about 60% of the population. This is comparing to Singapore’s first wave of the virus in April 2020 which has no people vaccinated. It is also observed that with higher number of people vaccinated, cases do not tend to spike as greatly as no people being vaccinated.

* 1. **Feature Engineering**

Only a handful of feature engineering was done to the dataset as the data is relatively clean. First, I selected the following columns:

|  |  |
| --- | --- |
| Column | Description |
| new\_cases | Number of new COVID-19 infection on that day |
| new\_deaths | Number of new deaths from COVID-19 infection on that day |
| people\_vaccinated | Number of people with the first dose of the vaccination in Singapore |
| people\_fully\_vaccinated | Number of people fully vaccinated in Singapore |
| stringency\_index | The Stringency of the measures put in place in Singapore |

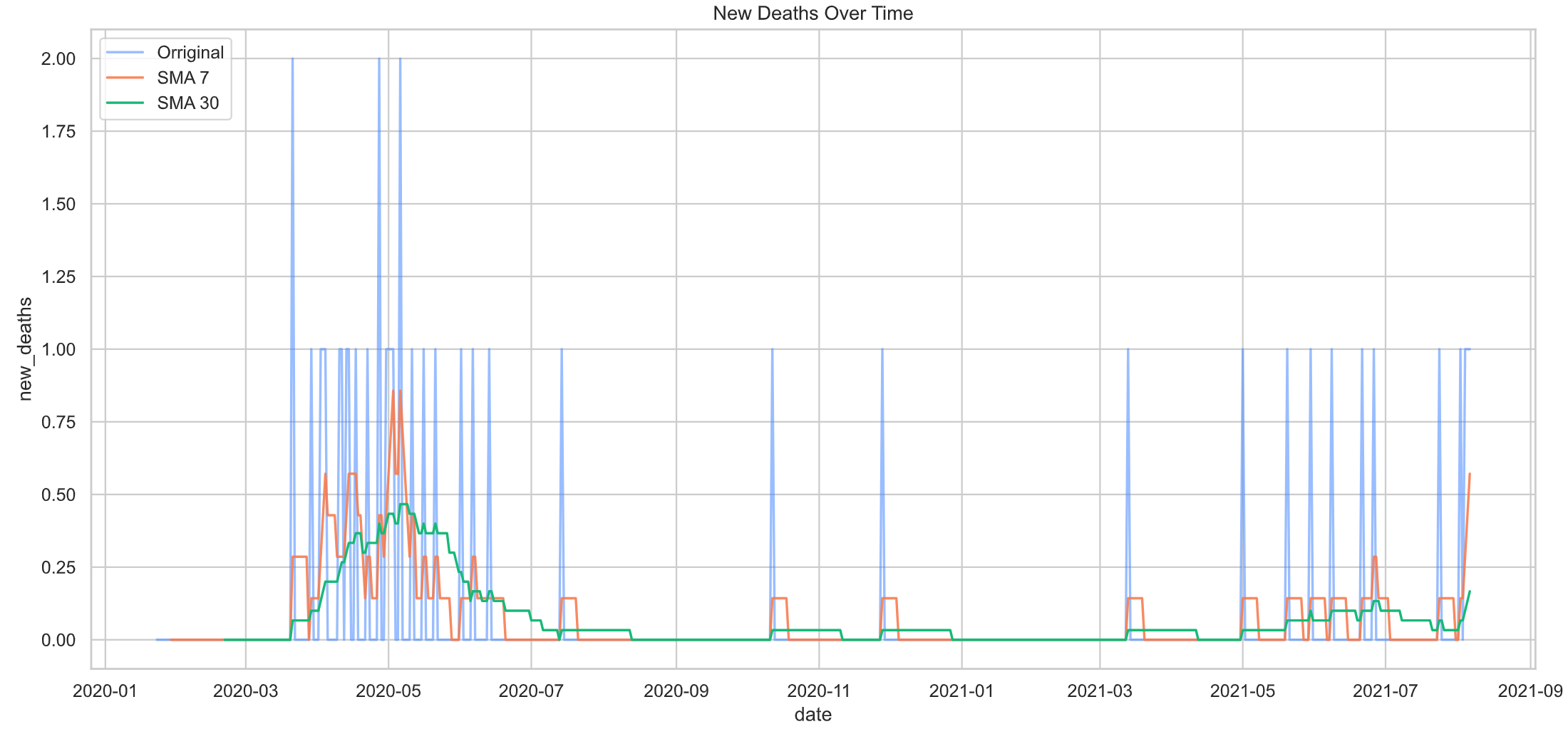
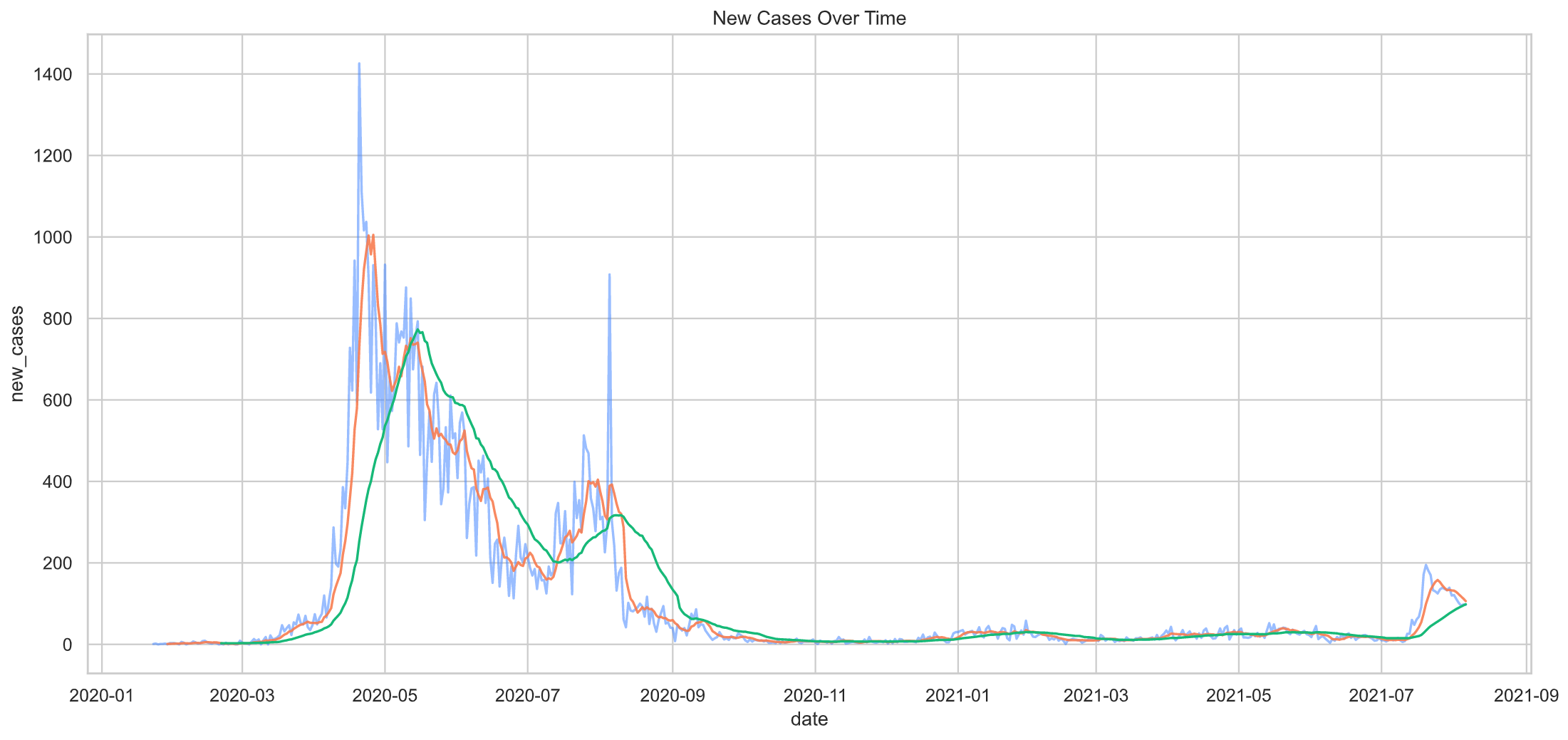
Next, all data before January 11 2021 will be fille with 0, following which, the .fillna method is applied with ffill. Finally, the data is split into training and testing with the last 28 rows used as testing.

1. **Methodology**

We will first be conducting various tests before training the VARMA model, under Section 4.3. After which, we will be comparing the VARMA model’s results with 2 ARIMA models. Despite the capability of the VARMA being able to predict more than one variable, the model does come with its own drawbacks. VARMA is unable to difference the dataset, a very important step to ensure our dataset is stationary. Differencing is required for this dataset which we will later see in Section 4.3.3. Therefore, comparing the 2 models will be interesting.

After all the tests and plots, we will be using TimeSeriesSplit with n\_splits=5 to evaluate the models. This would equate too about 3 months of validation data for each iteration. We will be looking at both the Training Root Mean Squared Error (RMSE), validation RMSE, Akaike’s Information Criteria (AIC) and means Bayesian Information Criteria (BIC). RMSE will be used to evaluate the accuracy of the model whilst AIC and BIC will be used to evaluate how well the model fits.

* 1. **Simple Moving Average (SMA)**

****

|  |  |
| --- | --- |
| Fig 6. SMA of New Cases Over Time | Fig 7. SMA of New Deaths Over Time |

We shall first take a look at the general trend of the data with the use of SMA. In figure 6, we can see that the 7-day SMA gives a better general trajectory of the number of new cases as compared to the 30-day SMA. Looking at the 7-day SMA, we can project that the number of new cases in the following days should be decreasing. In figure 7, the 7-day SMA also gives a better trajectory of the number of new deaths as compared to 30-day SMA. From the 7-day SMA, it is observed that the number of new deaths will continue to rise.

* 1. **Vector Auto Regressor Moving Average (VARMA) modeling**

Before we begin modeling, it would be good to understand some basics of the VARMA model. VARMA stands for Vector Auto Regressor Moving Average. Hence, there are 2 parameters for us to configure.

|  |  |
| --- | --- |
| p | Order of Auto Regressive Component |
| q | Order of Moving Average Component |

Due to limited time and computing power, we shall be finding out the order of the Auto Regressive (AR) component and Moving Average (MA) component manually with various plots and tests. On top of that, there are 2 basic requirements before we can use the VARMA model:

1. Need at least 2 variables
2. Variables should influence each other

As we are attempting to predict both number of new deaths and new cases, we have met the first requirement. However, we need to ensure we have met the second requirement.

* + 1. **Granger’s Causality Test**

VARMA works on the assumption that the variables influence each other. With the use of the Granger’s Causality Test, we can test this relationship before building the VARMA model. Assume we have two variables, Y1 and Y2. After testing for Granger’s Causality, the null and alternate hypothesis will be as such.

|  |  |
| --- | --- |
| H0 | Y2 does not Granger cause Y1 |
| H1 | Y2 Granger cause Y1 |

From the Granger test, we will get a P-Value. If the P-value is less than 0.05, we reject the null hypothesis (H0). To conduct the Granger’s Causality Test, we import grangercausalitytests from statsmodels.tsa.stattools. After conducting the test, we got the following results:

|  |  |
| --- | --- |
| Granger Test | P-Value (4 d.p.) |
| New Deaths Granger causes New Cases | 0.0 |
| New Cases causes New Deaths | 0.0 |

As both tests derived a P-Value of , we can reject the null hypothesis and conclude that New Cases Causes New Deaths and vice-versa.

* + 1. **Cointegration Test**

Cointegration test helps to test if two or more variables are significantly connected and have a long running relationship. Conducting the cointegration test gives us the test statistic and the critical value. If the test statistic is more than the critical value, the variables are said to have a long running relationship. Conducting the cointegration test is important as the VARMA model assumes that the variables are cointegrated. To conduct the Cointegration Test, we import coint\_johansen from statsmodels.vector\_ar.vecm. After conducting the test, we got the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Test Statistic | Critical Value (95%) | Significant |
| new\_deaths | 66.937 | 12.321 | True |
| new\_cases | 3.751 | 4.130 | False |

As new\_death Test Statistic is more than the 95% Critical Value, we can conclude that cointegration is present in new\_deaths. However, new\_cases Test Statistic is less than the 95% Critical Value, hence, we can conclude that cointegration is not present in new\_cases.

* + 1. **Stationary Test**

Ttime series models work on the assumption that the data is stationary. VARMA is no exception to this requirement. As a result, we shall be conducting the Augmented Dickey-Fully test. When conducting the Stationary test, the null and alternate hypothesis is set up as such:

|  |  |
| --- | --- |
| H0 | Variable is non-stationary |
| H1 | Variable is stationary |

To conduct the Augmented Dickey-Fuller test, we import adfuller from statemodes.tsa.stattools. After conducting the test, we got the following result:

|  |  |
| --- | --- |
| Variable | P-Value (95%) |
| new\_deaths | 0.115 |
| new\_cases | 0.250 |

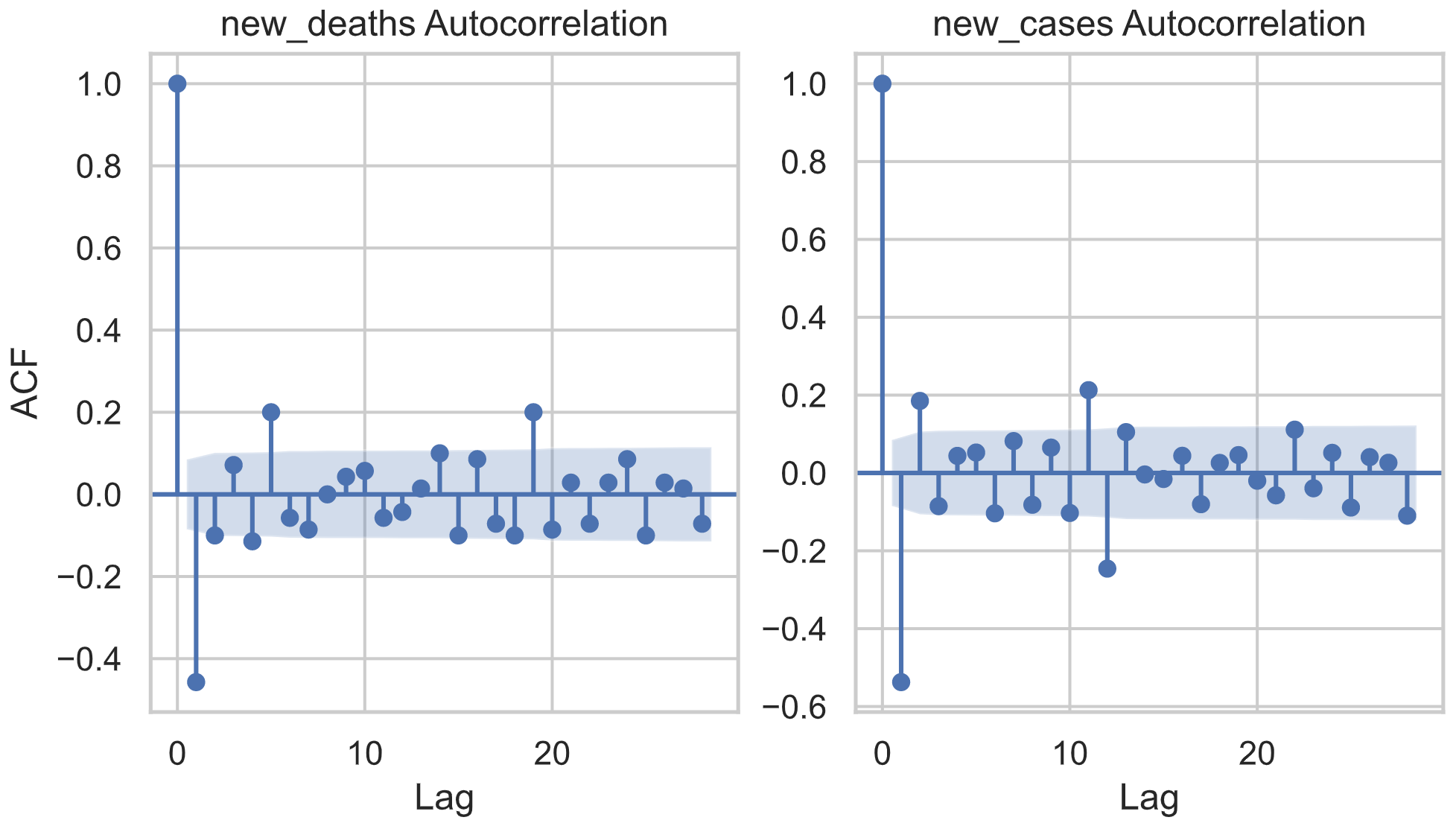
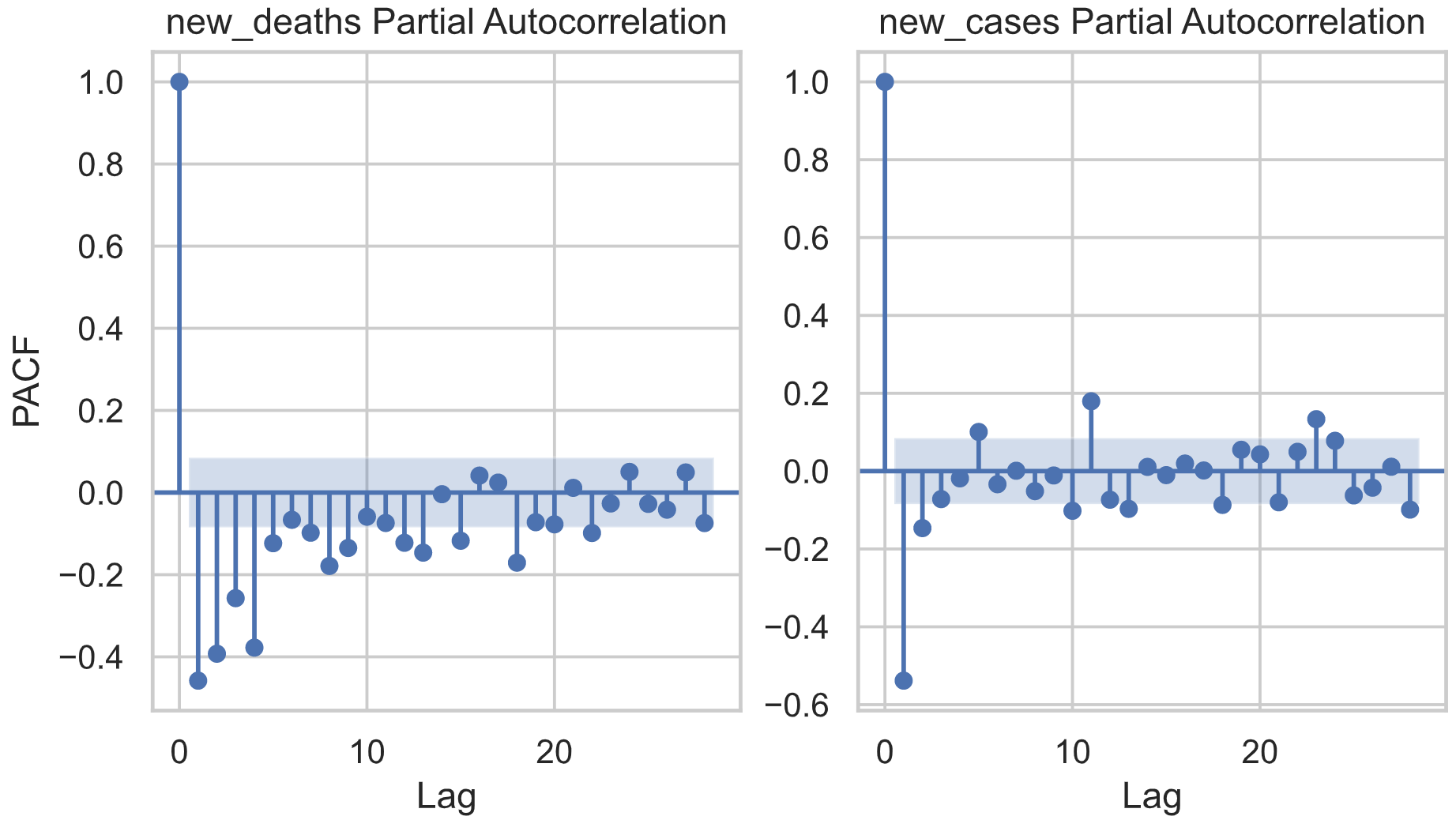
As both P-Values are more than 0.05, we cannot reject the null hypothesis (H0). As a result, we conclude that both variables are non-stationary. Since the dataset is non-stationary, we are required to difference the dataset. To difference the dataset, we simply run df.diff().dropna(). After differencing, the test was conducted again, and we got the following result:

|  |  |
| --- | --- |
| Variable | P-Value (95%) |
| new\_deaths | 0.0 |
| new\_cases | 0.0 |

As both P-Values are less than 0.05, we reject the null hypothesis (H0) and conclude that both columns are stationary after one round of differencing. Note, due to the limitations of the VARMA model, we are unable to apply differencing. Instead, in the ARIMA model we will be testing, the order of differencing (d) will be set to 1.

* + 1. **ACF and PACF Plots**

This section will allow us to determine the order of the AR (p) and the order of the MA (q). These values can be used for both the VARMA model and the ARIMA model. With the use of the plot\_acf and plot\_pacf functions under statsmodels.graphics.tsaplots. Plotting both plots for both columns respectively yields the following results:

|  |  |
| --- | --- |
| Fig 10. ACF for New Cases and New Deaths | Fig 11. PACF for New Cases and New Deaths |

In figure 10, both columns have a value of positive 1 on their first lag. The lags also show a gradual decrease towards 0. With this current understanding, we should stick to an AR model with order of MA set to 0. The ACF also decreases significantly after 2 lags.

In Figure 11, both columns also depict a gradual decrease towards 0. With this understanding, we will be using both the AR and MA model. It is also noticed that for the new\_cases column, PACF decreases significantly after 2 lags. As a result, we shall set the order of AR (p) to 2. Basing of Figure 10, we shall set the order of MA (q) to 2.

* + 1. **VARMA Performance Evaluation**

The VARMA model was first tested without the order, then with the order of (2, 2). Testing the model yield the following result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | Train RMSE | Validate RMSE | AIC | BIC |
| Without Order | 53.958 | 835.923 | 3416 | 3468 |
| With Order | 46.853 | 785.304 | 3368 | 3462 |

The model tested with the order not only yield a better RMSE, but it also had better fit based on the lower AIC and BIC score. However, the validation RMSE is very high and unsatisfactory. This could be due to the lack of differencing.

* 1. **Auto Regression Integrated Moving Average (ARIMA) Modelling**

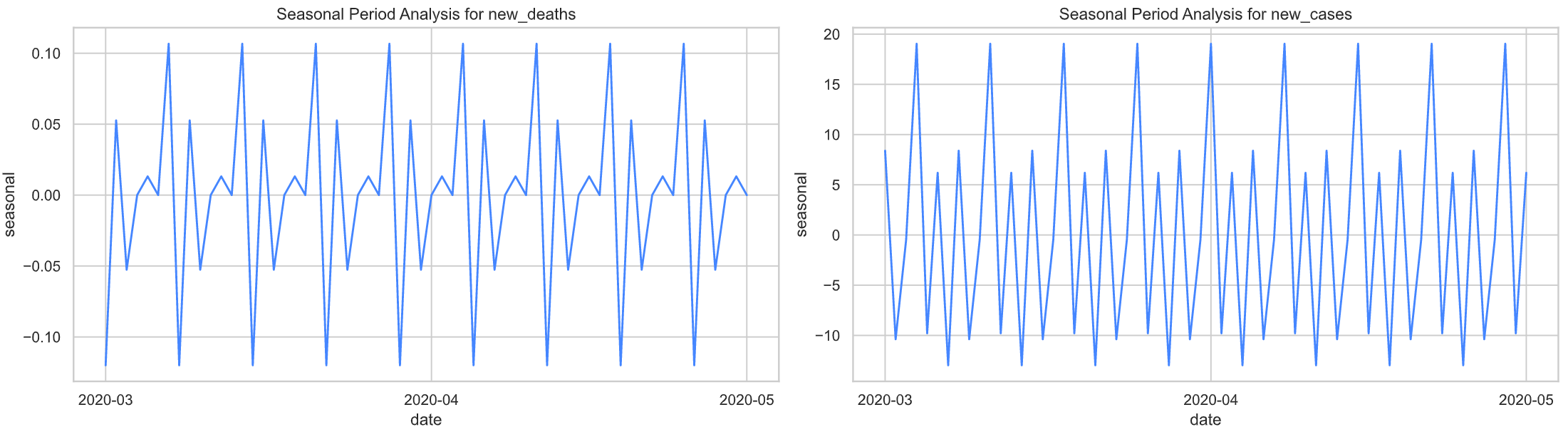
Before we begin modeling, it would be good to understand some basics of the ARIMA model. ARIMA stands for Auto Regressor Integrated Moving Average. Hence, there are 3 value to configure.

|  |  |
| --- | --- |
| p | Order of Auto Regressive Component |
| d | Amount of Differencing |
| q | Order of Moving Average Component |

In the ARIMA model, less tests are required. Only the Stationary Test is required which was carried out in Section 4.3.3. From which, we understand that we have to set the amount of differencing (d) to 1. We have also previously plotted ACF and PACF plots in Section 4.3.4. From which, we understand we have to set Order of AR (p) to 2 and Order of MA (q) to 2.

* + 1. **Seasonal Decomposition**

To create a more accurate time series model, we can use the seasonal order. Since the statsmodels library does not have any Seasonal VARMA model, this section is dedicated to the Seasonal ARIMA model as we can specify seasonal order in statsmodels ARIMA model. With the seasonal\_decompose function in statsmodels.tsa.seasonal, we parse the differenced training data. After which, we get an output containing the trend, seasonal and residual lots. Storing the seasonal output of each column, we then take a closer look at the seasonal pattern spanning 2 months.



|  |  |
| --- | --- |
| Fig 12. Seasonal Period Analysis for New Deaths | Fig 13. Seasonal Period Analysis for New Deaths |

From figure 12 and 13, it is observed that there are about 4 intervals in one month. This means each interval is about a week or 7 days. Based off this understanding, we should set number of period (m) to 7. Next, we carry on to the Seasonality Stationary Test.

* + 1. **Seasonality Stationary Test**

The Seasonality Stationary Test determines if the seasonal data is stationary. If our Seasonal data is not stationary, we will need to apply seasonal differencing. When conducting the Stationary test, the null and alternate hypothesis is set up as such:

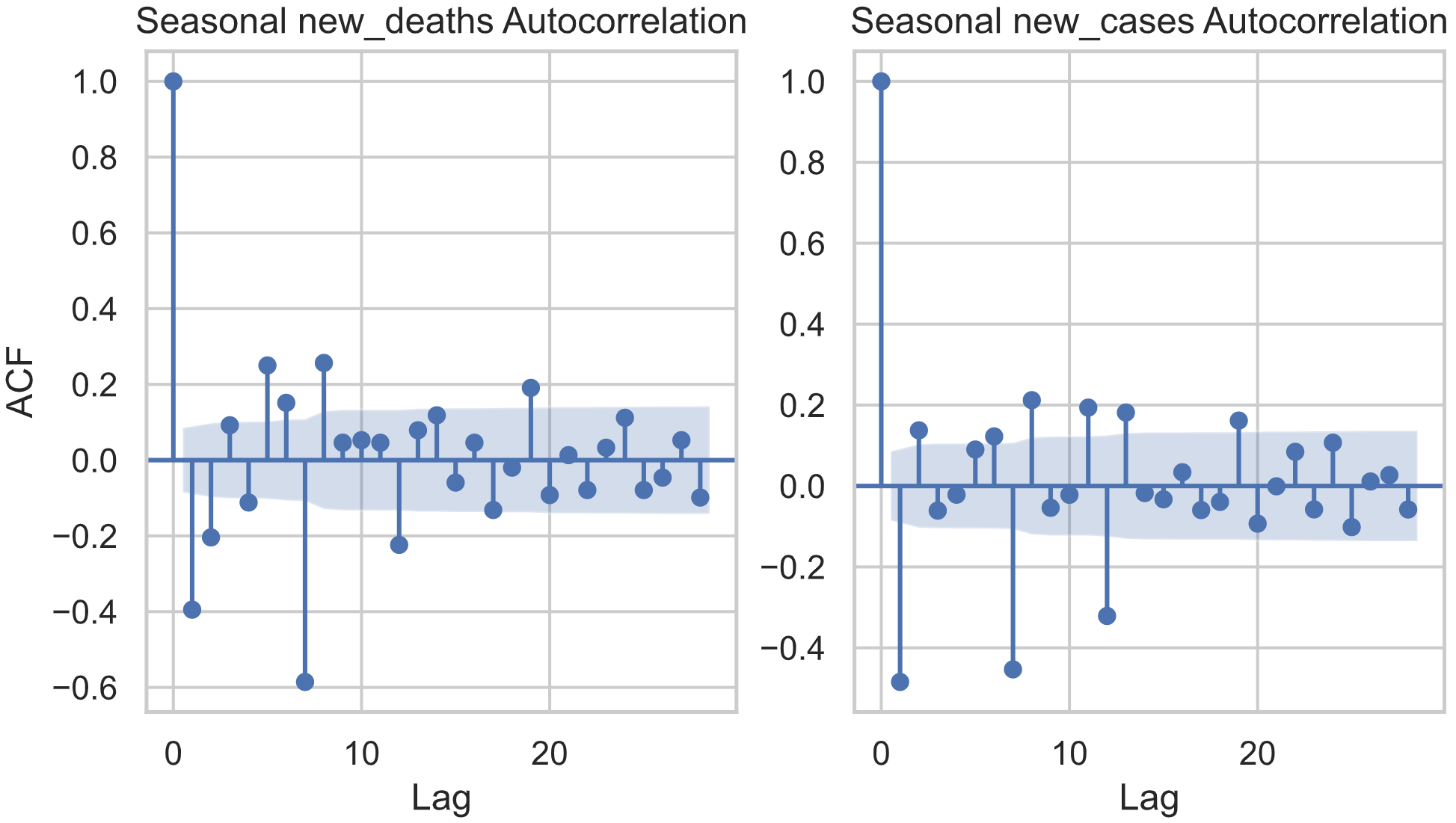
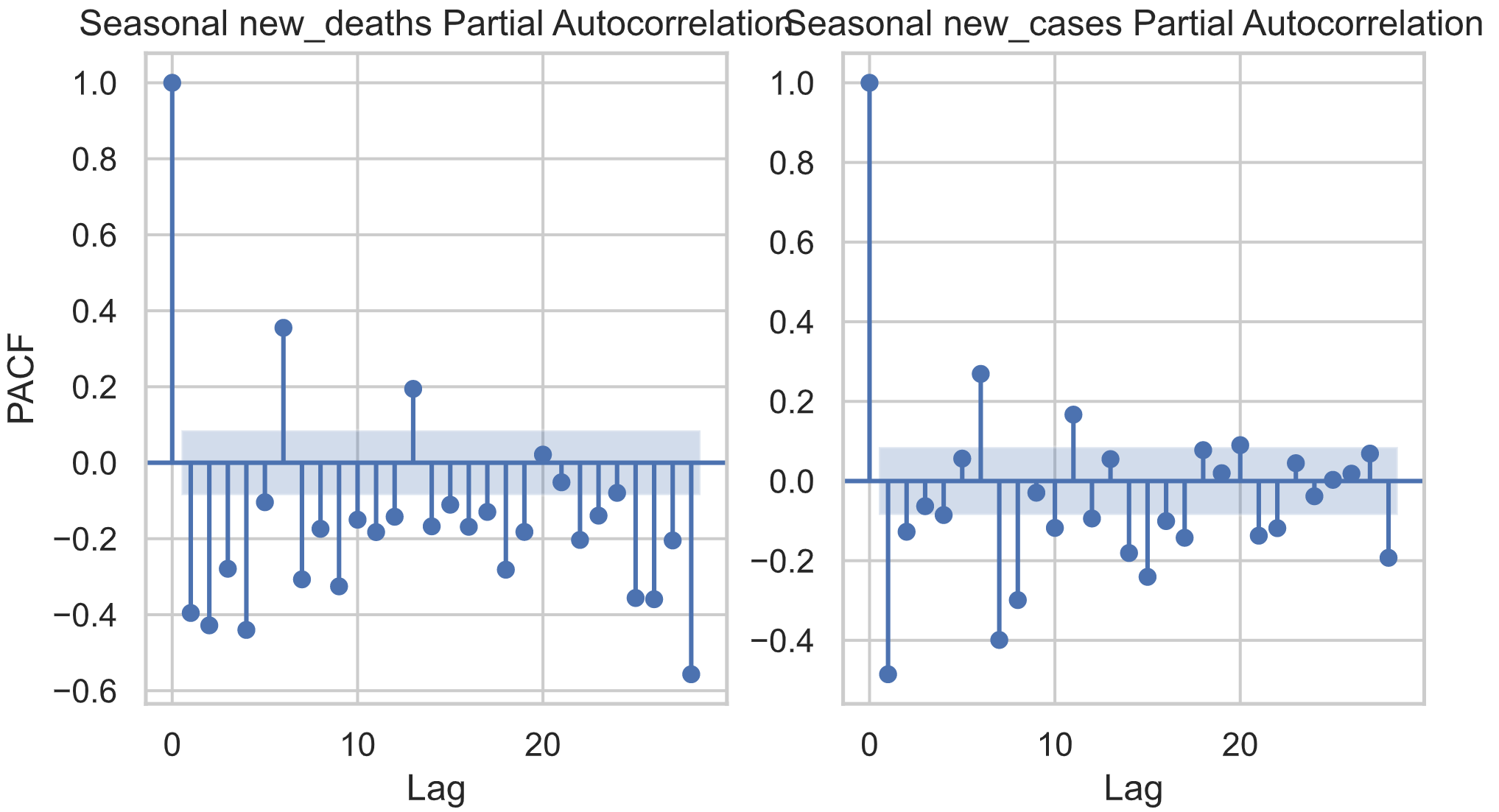
|  |  |
| --- | --- |
| H0 | Variable is non-stationary |
| H1 | Variable is stationary |

To conduct the Augmented Dickey-Fuller test, we import adfuller from statemodes.tsa.stattools. After conducting the test, we got the following result:

|  |  |
| --- | --- |
| Variable | P-Value (95%) |
| new\_deaths | 0.0 |
| new\_cases | 0.0 |

As both P-Values are less than 0.05, we reject the null hypothesis (H0). We conclude that both Seasonal variables are stationary and no Seasonal Differencing is needed. As a result, amount Seasonal Differencing (D) will be set to 0.

* + 1. **Seasonal ACF and PACF plots**

|  |  |
| --- | --- |
| Fig 14. Seasonal ACF for New Cases and New Deaths | Fig 15. Seasonal PACF for New Cases and New Deaths |

In Figure 14, the first lag has a value of positive 1. The lags also show a gradual decrease towards 0. With this current understanding, we would stick to a Seasonal AR (P) model with Seasonal MA (Q) set to 0.

In Figure 15, both variables also depict a gradual decrease towards 0. With this understanding, we will be using both the AR and MA model. It is also noticed that for the new\_cases column, PACF decreases significantly after 2 lags. As a result, we shall set the order of AR (p) to 2. Basing of Figure 14, we shall set the order of MA (q) to 2.

* + 1. **SARIMA Performance Evaluation**

The SARIMA model was first tested without order or seasonal order. Then it is tested with the order followed by order and seasonal order. After testing, the results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | Train RMSE | Validate RMSE | AIC | BIC |
| Without Order | 78.936 | 885.431 | 2270 | 2287 |
| With Order ONLY | 46.864 | 167.098 | 1670 | 1698 |
| With Order and Seasonal Order | 46.369 | 122.414 | 1672 | 1714 |

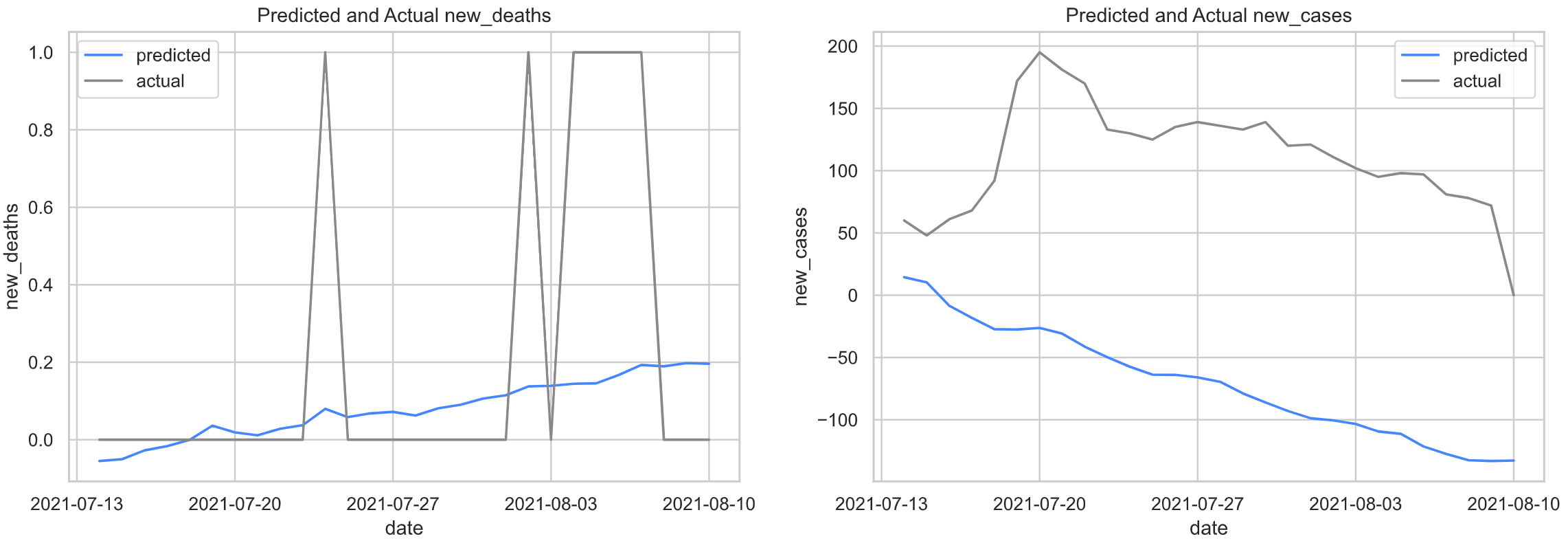
The ARIMA model without any Order or Seasonal Order performed the worst. When ARIMA was tested with the Order, the Training and Validation RMSE dropped significantly to. The RMSE further dropped when the Seasonal Order was included. However, the model with both Order and Seasonal order has a slightly bad fit when compared to the Model with Order only based of the AIC and BIC.

1. **Model Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Model Type | Train RMSE | Validate RMSE | AIC | BIC |
| VARMA | Without Order | 53.958 | 835.923 | 3416 | 3468 |
| VARMA | With Order | 46.853 | 785.304 | 3368 | 3462 |
| ARIMA | Without Order | 78.936 | 885.431 | 2270 | 2287 |
| ARIMA | With Order ONLY | 46.864 | 167.098 | 1670 | 1698 |
| ARIMA | With Order and Seasonal Order | 46.369 | 122.414 | 1672 | 1714 |

When placing the results of the VARMA and ARIMA models side by side, we can see that the ARIMA model outperforms the VARMA model. Although the VARMA model performs better then the ARIMA model when order is not specified, with order specified, the ARIMA truly shines. On top of that, all 3 ARIMA models tested have better fits based on the AIC and BIC. Another thing to take note about the VARMA model is, it took significantly longer to train as compared to the ARIMA model.

Based on the Validation RMSE, we should select the ARIMA model with both Order and Seasonal Order. When testing the final model with the COVID-19 testing set, results were satisfactory.



|  |  |
| --- | --- |
| Fig 16. Predicted and Actual New Deaths | Fig 17. Predicted and Actual New Cases |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Train RMSE | Validate RMSE | AIC | BIC |
| new\_deaths | 0.258 | 0.405 | 94 | 150 |
| new\_cases | 79.488 | 188.305 | 6528 | 6314 |

From the graph, we can see that the model is able to capture the general trend for both variables. However, it is unable to accurately predict in detail. For example, it is unable to predict the sudden rise of cases before the drop of cases. This could largely be because, the cases before 13th of July were stagnant.

1. **Discussion**

Forecasting Singapore COVID-19 cases and deaths is a possibility. Although the forecast might not be accurate. It is important to note that a decent forecast is better than no forecast. A forecast can better help us determine when restrictions should be put in place and warn the public before the COVID-19 situation gets out of hand.

Basing of this paper’s ARIMA model however, we should only use the forecast as a recommendation of when measures should be put in place. This is because the results lack reliability in the prediction of COVID-19 cases and deaths. However, better models can be used for the forecasting instead of ARIMA. In fact, what Singapore National Center for Infectious Diseases (NCID) or Ministry of Health (MOH) can do is, host a competition to forecast COVID-19 infections. After which, select the best 20 models and get the average prediction from these 20 models. This would greatly improve the accuracy of the prediction as proven by the CDC’s COVID-19 forecast.

For a more accurate prediction, forecast can be done every day, and more metrics could be used such as number of people tested. These features selected are within control of MOH. This is why features such as stringency and vaccination rate were used in this paper as these features can be confidently forecast.

1. **Conclusion**

Overall, it has been very insightful to work on this technical paper. The main learning point was time series forecasting of multi-variate data with the use of the VARMA model. We also learnt that multiple SARIMA model performs better compared to a single VARMA model in accuracy, fit and speed.

It was also quite interesting to work on something that is very relatable and understand that time series forecasting is a difficult but useful endeavor.

1. **References**

Audrey Tan, 2020, ‘Study using AI finds more effective drug mix to fight coronavirus’, *The Straits Times*, May 30, accessed 11/8/2021

<<https://www.straitstimes.com/singapore/health/study-using-ai-finds-more-effective-drug-mix-to-fight-virus>>

CDC, 2021, ‘COVID-19 Forecasts: Cases’, *Today*, August 11, accessed 11/8/2021

<<https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html>>

Deepak Painuli, 2021, ‘Forecast and prediction of COVID-19 using machine learning’ *National Center for Biotechnology Information*, May 21, accessed 11/8/2021

<<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/>>

Department of Statistics Singapore, 2021, ‘Singapore Population’, June 30, accessed 11/8/2021

<<https://www.singstat.gov.sg/modules/infographics/population>>

Jennifer Chu, 2020, ‘Artificial intelligence model detects asymptomatic Covid-19 infections through cellphone-recorded coughs’ *MIT News*, October 29, accessed 11/8/2021

<<https://news.mit.edu/2020/covid-19-cough-cellphone-detection-1029>>

Michael Yong, 2020, ‘Timeline: How the COVID-19 outbreak has evolved in Singapore so far’, *Channel News Asia*, April 18, accessed 11/8/2021

<<https://www.channelnewsasia.com/singapore/singapore-covid-19-outbreak-evolved-coronavirus-deaths-timeline-764126>>

Youyan Gu, 2020, ‘About covid19-projections.com’ April 30, accessed 11/8/2021

<<https://covid19-projections.com/about/>>