MATH334 Group Project: COVID Cases and Vaccinations in India and Italy

T. Stockton*, H. Horton†, A. Williams‡ and B. Mullally§

*37552384
†34797769
‡37607332
§37545671

***Abstract*—This report researches into 4 separate time series on total cases and vaccinations for India and Italy during the COVID-19 pandemic. ARIMA(0,2,3), ARIMA(0,2,1), ARIMA(13,2,0) and ARIMA(3,1,0) models were found using the Box-Jenkins method and were used to forecast the future of the respected time series.**

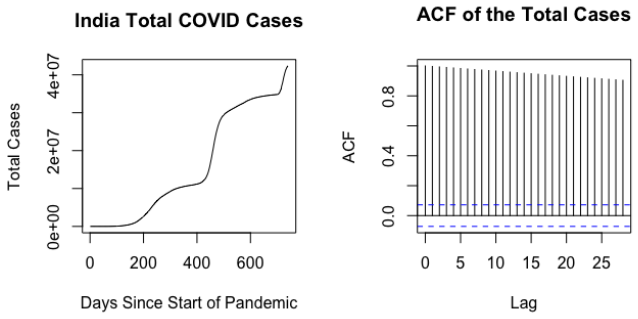
I. Introduction

In this report we are going to investigate and model time series' of India and Italy. Looking at their total coronavirus cases and their vaccination rates. We have chose these two countries as they were both highly affected by the virus, with Italy being the first European country to enter a lockdown at the very start of the pandemic. Meanwhile, India reported around 400,000 cases a day in May 2021, and was at one point responsible for more than half the worlds daily COVID cases [1]. Time-series analysis is important when looking at COVID data as it could be, and is used by governments to help forecast what may happen in the future based on the current data. This helps them decide whether to introduce/remove extra measures that may be needed to help reduce the spread of COVID. We chose total cases and vaccinations so that the forecasts could be used to make sure that the countries have enough vaccines and we can see the current vaccination rates and compare with the cases predictions to see if enough is being done to combat the spread of the virus.

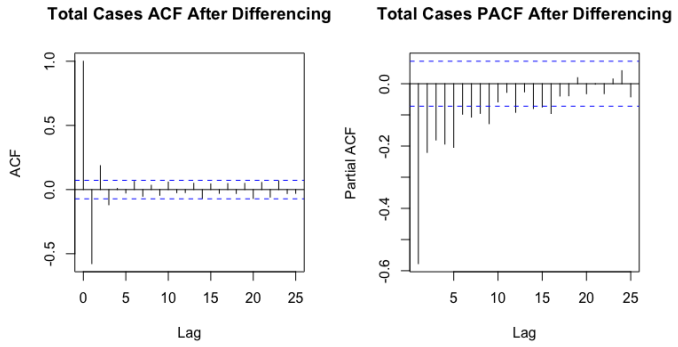
II. India

A. Total Cases

First we will look at the India Total Cases time series:



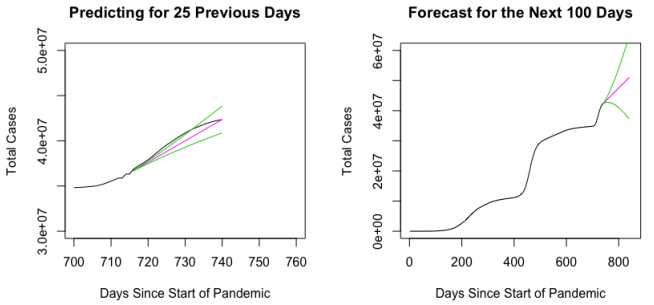
Here we see in the time series we have two main large spikes, one just after 400 days since the start of the pandemic, and the other at roughly 700 days. These two spikes correlate to the rise of the two COVID variants: the delta variant; and the omicron variant. Both of which have had massive impacts on COVID cases within India, especially the delta variant which was first identified within India. From the ACF above, we see a very slow decay suggesting that the data is not stationary. This implies we need to apply differencing to the time series in order to make it stationary. First we logged the time series to stabilise the variance. Then we differenced it, however upon applying a Dickey Fuller (DF) test we got the p-value, $p = 0.1557$ which is greater than our significance level of 0.05. This means the data is still not stationary so we have to difference the data again to get a DF test value < 0.05. After two differences we found our data was stationary so now we could predict a model for the data. So we now plotted our ACF and Partial ACF (PACF) in order to find a model for our data:



Using this ACF and PACF we can estimate that the time series can either be modelled by a Moving Average, MA(3) or an Auto-regressive, AR(16) model since these are the final significant values in the ACF and PACF respectively. Since, an MA(3) model is considerably less complex than an AR(16) model, we will use this model as long as it passes the Ljung-Box test when applied to the data. When modelling the data by this MA(3) model we get that the mean of the residuals is 0.000219, suggesting the model is a good fit, despite the first 100 or so values having slightly large residuals due to the early days having very small amounts of cases. To improve this we could have removed the first 100 values of this time series, however we didn't feel it was necessary as the mean

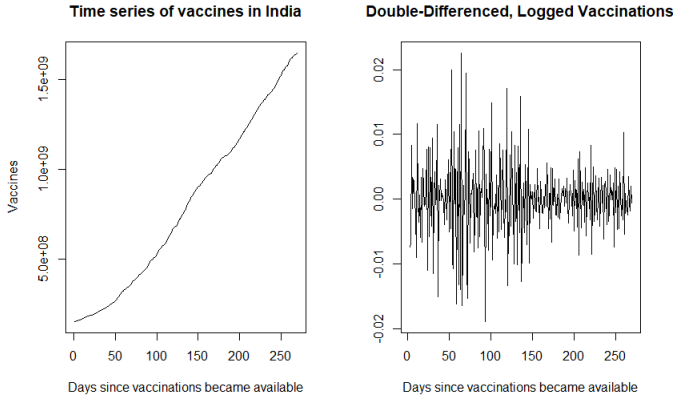
of residuals was already so close to 0. Then we performed a Ljung-Box Test to test whether the residuals are distributed by an i.i.d white noise process. Upon performing a Ljung-Box test with lag=28 and degrees of freedom=3, we got the p-value, $p = 0.07$ which is > 0.05 so there is insufficient evidence to reject the null hypothesis. Therefore this model is suitable to use to model this time series.

We now have a suitable model, which is an MA(3) that is differenced twice, so we now have an ARIMA(0,2,3) model. We will use this to predict and forecast on the actual time series. Here we see a prediction of the the previous 25 days using the model, as well as a forecast for the next 100 days:

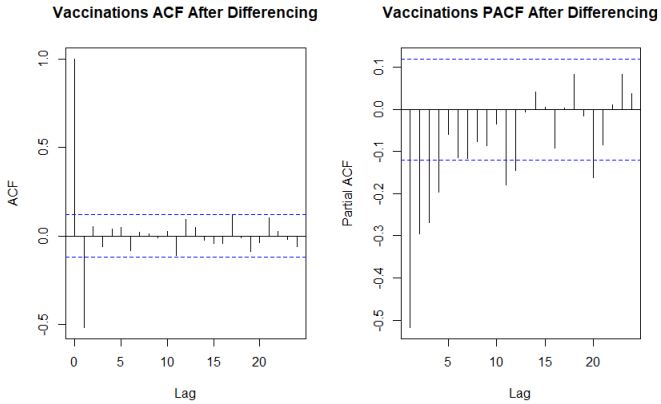


In these graphs the black line represents the total cases time series, the pink line represents the prediction/forecast, and the green line represents the 95% confidence interval for the prediction/forecast. In the prediction for the previous 25 days we see it seems relatively accurate compared to the actual time series. We have a Mean Absolute Error (MAE) of 529,750.6 and a Root Mean Squared Error (RMSE) of 593,029.2, which are both relatively small in context to the large values in the time series. However, there is an issue that arises. This is that the actual time series does leave the 95% confidence interval, which may be due to the model trying to forecast whilst a spike is occurring, which it could not predict fully. However, the total cases quickly returns to being similar to the prediction, and by the end of the prediction we see they are almost identical in value. The forecast for the next 100 days suggests a steady increase in COVID cases in India, with no spikes or plateaus. However, the confidence intervals are very large suggesting there's a good chance the actual total cases may vary considerably compared to the forecast. This is likely due to the fact it is almost impossible to predict when or how contagious a new variant may be. Governments could still use these time series alongside further knowledge of a new variant emerging to create forecasts of the effects of the new variant. They could use this time series to help model a forecast by using the data on the previous spikes. We personally think the forecast may have been slightly more accurate if it did show signs of plateau as that is what we see tends to happen after a spike. This could be down to either: the government introducing stricter measures to slow the spread; or as we have seen with Omicron in South Africa, it naturally plateauing as the majority of the population have been infected.

B. Total Vaccinations



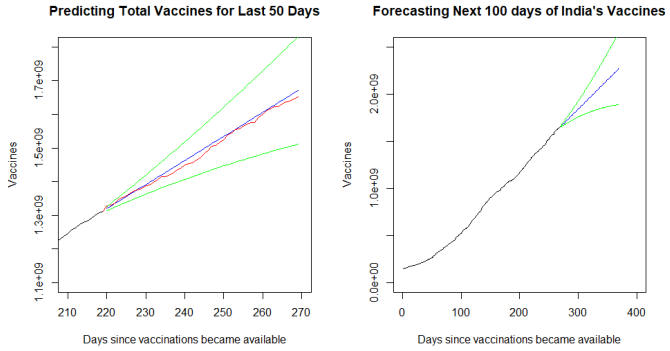
Here we see the time series of India's total vaccinations after the removal of the first 400 days of data. The first 300 days did not have any recorded vaccinations as no vaccine existed but the next 100 days were chosen to be removed as vaccination rates were slow which massively affected the residuals, making model finding very difficult later on. As you can see from the graph there is a steady, linear increase in total vaccinations but it is clearly non-stationary. Similar to India's total cases time series, the ACF of the vaccinations is gradually decreasing which indicates that differencing may need to be done. Again, we logged the data to stabilise the variance. A DF test was done which produced a very large p-value of 0.9767 meaning differencing must be done. After one difference the DF p-value reduced to 0.2464. We differenced again and the DF p-value then reduced to below 0.01. From the graph, you can clearly see the new time series is stationary which backs up the hypothesis behind the DF value.



The ACF and PACF of the transformed time series above show that there is clear potential for many models for this data. An AR(4) or AR(11) model could have potential for example with the clear spikes but there are a lot of possible models to try. We decided to use an MA(1) model however due to it being the only spike in ACF and it being the simplest model.

The mean of the residuals of the MA(1) model was -0.00005 and the Ljung-Box p-values are high even for very

large lags therefore there is sufficient evidence to accept our null hypothesis and so the residuals are an i.i.d. white noise process (see appendix A). As a result, this model is suitable and is likely to give good forecasts.



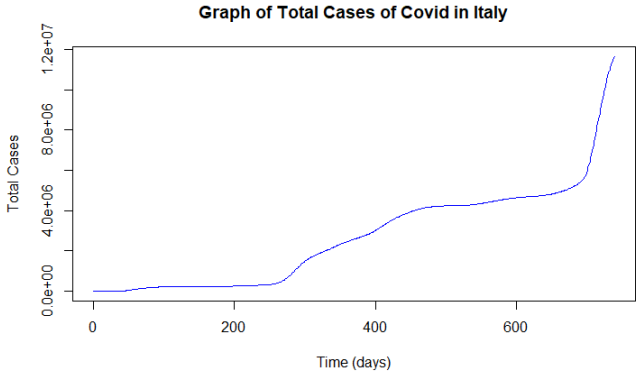
By removing the final 50 days from the original data, we have created a training set and used this to predict the missing values along with the ARIMA(0,2,1) model found earlier. The blue line shows the prediction along with the actual data in red and the 95% confidence region bounds in green. The prediction looks extremely accurate from this graphical representation. The MAE of the prediction was 20,857,926 and the RMSE was 49,072,082. Although these values are large, we are handling massive numbers in the millions and the graph shows the prediction is still very accurate. Finally, we forecast 100 days into the future to show what India's total vaccinations could look like (with the error bounds in green). It shows that they are likely to continue on the same linear trajectory which is very likely however obviously there is a limit to the people that can be vaccinated due to the population and so the graph will eventually plateau.

From these two time series' so far we notice that the introduction of vaccinations and their rise in numbers seems to correlate with the decline in the spike of total caused by the Delta variant (the spike soon after day 400). This suggests the vaccinations are effective in the preventing the spread of the Delta variant, however they do not seem to have such effectiveness on the. Omicron variant that still. managed to spike despite vaccine numbers being higher than before. We do. however see that the Omicron variant seems to be a much smaller spike suggesting the vaccine is having some influence on infections caused by the Omicron variant in India.

III. Italy

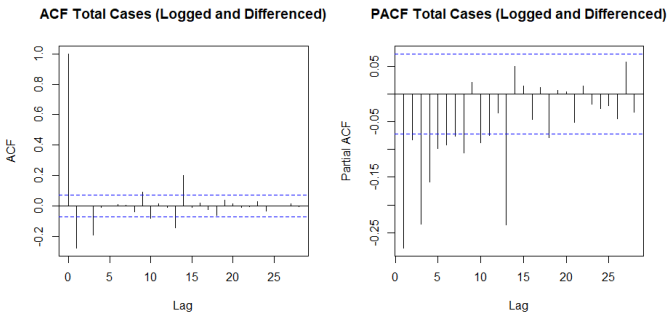
A. Total Cases

Now we will look at the Italy Total cases time series:



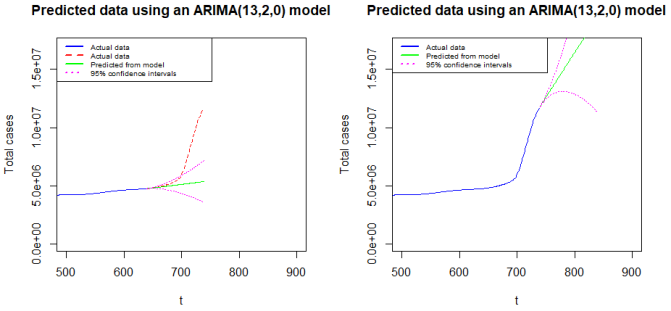
From above, we can see two significant spikes, one at about 250 days after the start of the pandemic and the other at 700 days. These two spikes correspond with the first being the initial wave of coronavirus in Italy, causing the lockdown[2], and the second is due to the Omicron variant. Italy was the first European country to undergo a major outbreak, and the first to instigate a national lockdown.

The ACF of our time series indicated that there is no sudden drop off as the lag increases, only a slight decrease over higher lag. This suggests that this time series is not stationary, so we must transform some of the data to make it stationary. As we only have positive values of data, we can log the data to stabilise the variance. We performed a Dickey-Fuller test to this logged data to determine if we needed to difference. Upon two DF tests with one being on the logged data and the other on the logged differenced data, we find our p-value to be above our significance level of $\alpha = 0.05$. This indicates that we need to difference at least 2 times, and on a third DF test we find that the p-value is significant thus finding our time series stationary. Now that we have stationary data, we can now look at our ACF and PACF of our logged differenced data to find a suitable model.



From the ACF and the PACF we can choose a model that best suits the data. Choosing either a Moving Average, $MA(14)$ or an Auto-regressive, $AR(13)$, based on the latest significant drop-offs of both ACF and PACF graphs respectively. As an $AR(13)$ model is less complex than a $MA(14)$ model. Before we say this will be our model, we need to check that the residuals follow an i.i.d. white noise process. For this we can perform the Ljung-Box test obtaining a p-value of 0.211, implying that there is no evidence to reject the null hypothesis that the residuals are white noise.

For more validation we can also look at the mean of the residuals which is -0.00000551. To improve this, we could have removed the first 100 values of the time series as it has quite a large variance for the first 100 residuals. However, as the mean is so close to zero, we thought this not necessary. The resulting suitable model is an $AR(13)$ model that has been differenced twice, $ARIMA(13, 2, 0)$. Now that we have a model, we will test by training it on a portion of the data and forecasting it against already known data. We will also plot the forecast for the next 100 days of after the data already known.

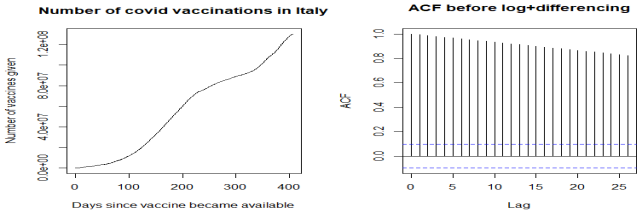


In the prediction for the previous 100 days, we see it seems relatively accurate compared to the actual time series. We have an MAE of 1598455 and an RMSE of 2630132, both being rather large. Additionally, the predicted time series also leaves the 95% confidence interval quite early at about 50 days and only predicts the next 25 days approximately correct. Based on the prediction for the last 100 days, when looking at the forecast for the next 100 days it would not be safe to assume that the prediction is very accurate. However, if we look at only the next 25 days, we can say that this model performs quite well on short-term predictions. It may be considered that an alternative model could be used such as an ARCH model, due to its ability to capture volatility that changes.

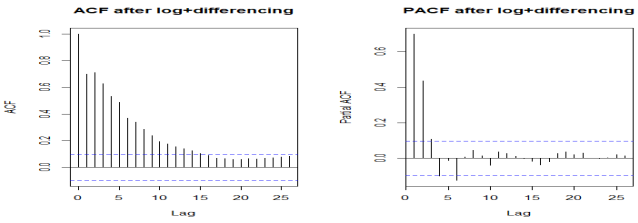
The total cases in Italy forecast compared to the total cases in India suggests that Italy are to be worse affected by the Omicron variant than India with a much steeper gradient of line in the forecast.

B. Total Vaccinations

Next we analysed the total vaccinations in Italy. The vaccine wasn't available in Italy until the 27^{th} December 2020[3] roughly 331 days after our COVID data began. We omitted these 331 days from the time series as no vaccine was available so they were all 0 values; this is so they did not have any affect on any model we come to use.



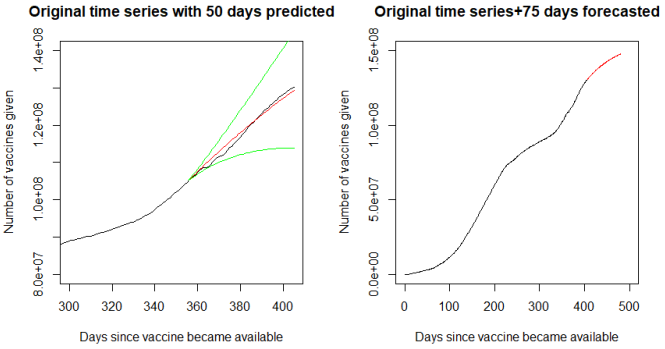
Similar to the other time series we logged the data then differenced it to make the data more stationary. Performing a dickey fuller test on this, with the null hypothesis that the data isn't stationary gives us a low p value meaning we can reject this null hypothesis and say the data is stationary. Below you can see the ACF and the PACF of the time series once it had been logged and differenced.



The gradual decrease of the ACF and the sharp drop offs at lag $2, 3$ on the PACF suggests an AR model would be best for this data. After trying both we decided to continue using an $AR(3)$ model, since the $AR(2)$ has very low Ljung-Box p values.

This is a very good model since the mean of the residuals is almost $0(-0.00093)$ as well as no significant spikes in the ACF or PACF, along with high P values for the Ljung-Box test across a range of lags.

Now to see if this model is useful for forecasting we will see how good it is at predicting the data we already have. To do this we took off the last 50 values from the original time series and then used my $AR(3)$ model to predict those same fifty values.



As you can see the prediction(shown in red) is not very far off the actual data(black), this prediction has a low MAE = 880,000 in the context of the numbers we are currently working with(120 million). This means any forecast will be reasonably reliable and could be used to give a rough estimate of how many more vaccines Italy should be expected to give out in the future for the reasons stated previously. Also next to it is a forecast for the next 75 days based of my $AR(3)$ model.We can see from the forecast that the number of vaccines being given will continue to increase for the next 75 days, but the rate at which it is increasing will start to slow down.

REFERENCES

[1] The New York Times (2021, November 17). *What to Know About India's Coronavirus*, Accessed: 16 Feb. 2022. [Online]. Available: https://www.nytimes.com/article/india-coronavirus-cases-deaths.html

[2] The Economist (2020, May 9). *Italy, the first country in Europe to enter lockdown, starts to emerge*, Accessed: 16 Feb.2022. [Online]. Available: https://www.economist.com/europe/2020/05/09/italy-the-first-country-in-europe-to-enter-lockdown-starts-to-emerge

[3] Wikipedia (2022, Jan 20). *COVID-19 vaccination it Italy*, Accessed: 16 Feb. 2022. [Online]. Available:https://en.wikipedia.org/wiki/COVID-19_vaccination_in_Italy: :text=The%20COVID%2D19%20vaccination%20campaign,countries%20in%20the%20European%20Union.

IV. APPENDIX

A.