334 Group report

HongSheng Lin 37527355, Michael Jennings 34374736

February 2021

Abstract

The report mainly included four parts .
1. The methodology of Box-Jenkins approach: Fit the COVID-19 data into time-series models and make reasonable forecast of cases/deaths of COVID-19.
2. Evaluation on forecast: Will there be a different on result of forecasting if start forecasting on different leading times and the polynomial trend estimation.
3. The seasonality of reported cases and regions: comparison between SARIMA and ARIMA .
4.Comparing models across region: the similarities and differences of models between each regions

1 Introduction

Time series analysis on the data of cases and deaths change all over U.S. would help the government or the public learn more about COVID-19. Make a proper forecast for reference.

2 Methods

The Box-Jenkins approach was been used to identify and estimate the model. First, Convert the COVID-19 data set into time-series then checking stationarity and Seasonality by Augmented Dickey- Fuller Test and auto correlation plot/ partial auto correlatiaon plot, a stationarity time-series would not have a very slow decay on auto correlation function plot. The method differencing will be used if the series is non-stationary. The seasonality will be detected by auto correlation function of the graph, the repeated pattern on fixed lags will be detected by auto correlation function plot and partial auto correlation if the seasonality exist . Second, to identify the model and p,d,q from ARIMA model by the pattern of auto correlation function and partial auto correlation function without seasonal effect. Then choose the proper model by examine the Akaike information criterion.

Move into the data now, The data we've given is a csv file with the cases and deaths of all states in U.S. from 2020/01/21 to 2021/01/09, First, extract the data of cases and deaths in California from 2020/01/21 to 2021/01/09. The original data of cases/deaths shown below.

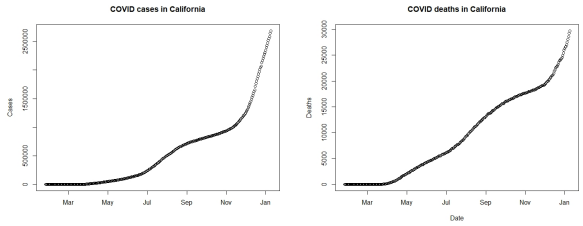

Figure 1: Plot of the original data of cases/deaths in California

modelling a pandemic often requires a log transformation since it's easy to assume that the cases are going to grow exponentially. Then, the series became stationary after log and take twice difference , Shown Figure 2, also to examine the p-value of adf.test to check stationary.The result shows stationarity is confirmed since we reject the null hypothesis of non-stationarity.

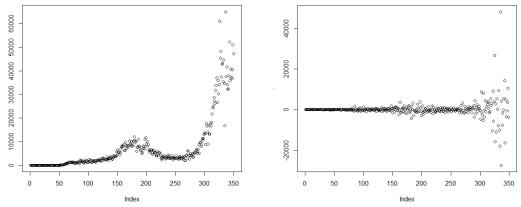

Figure 2: Plot of the difference once and twice data of cases/deaths in California

Then,to look into the auto correlation function and partial auto correlation function to identify the value of p,d,q in ARIMA ,examine(4,0,0)(1,0,0)(2,0,0), Also,after second differencing look stationary p-value less than 0.05. Time-plots and assessment of the ACF of the differenced series indicates that there may be a seasonal periodicity. In both situations the ACF peaks every 7 lags. To model this seasonal behaviour, we can try to take seasonal differences. After a few attempts, the diagnostics of mod SARIMA (3,2,0,1,1,1) is more suitable for the data of cases in california ,and the model SARIMA (5,2,0,1,1,1)is suitable for the deaths data in California.Because both of them owns relative high p-values out to large lags after Ljung-Box test and the residual checking .Also, the predictions were been made in the below Results part. After, apply the

methodology of Box-Jenkins approach for different starting week to research the problem of rolling start data for forecasting.

3 Results

1.Final models after calculating are SARIMA(3,2,0,1,1,1) for cases in California and SARIMA(5,2,0,1,1,1)for deaths in California the graph with prediction shown by Fig.3(lowest width of resiual .
2.Different starting week do make difference on prediction.Shown by fig 4 and 5 compare to Fig3.
The seasonal model is more reasonable than non-seasonal model according to the prediction. Shown by Fig.6

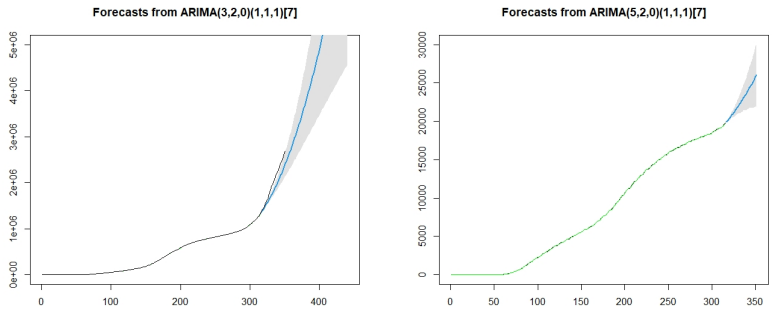

Figure 3: SARIMA(3,2,0,1,1,1) and SARIMA(5,2,0,1,1,1) with prediction start from week1.

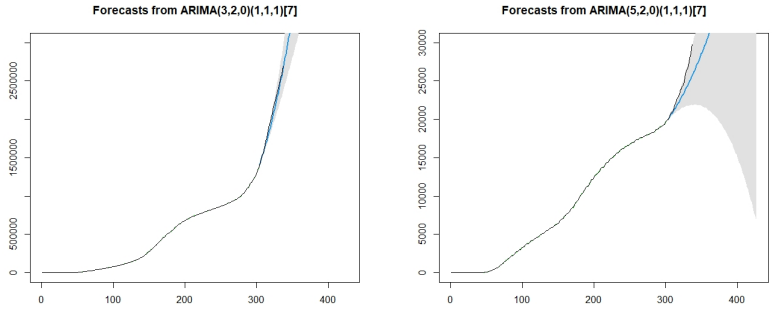

Figure 4: Prediction start from week2.

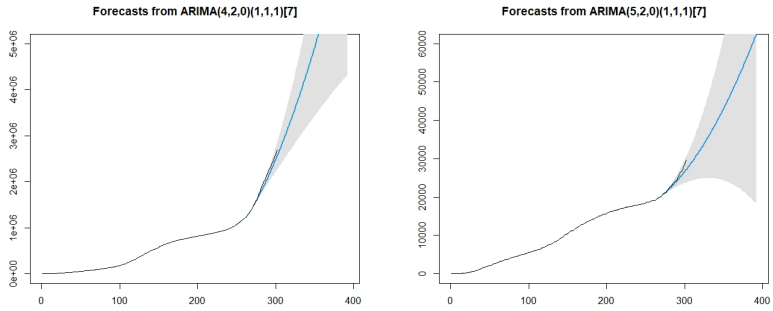

Figure 5: Prediction starts from week5.

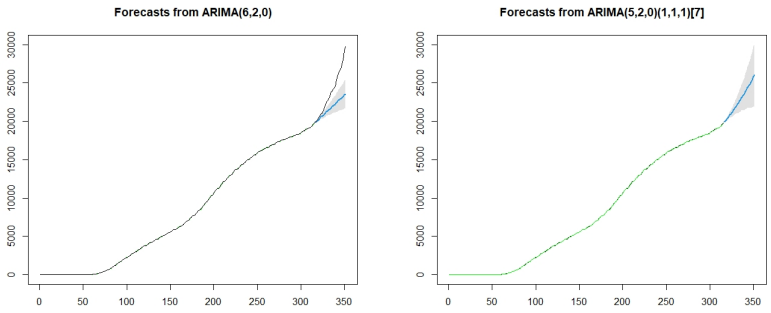

Figure 6: Modelling with/without seasonal effect.

3.1 Quadratic trend

After analysing the results from the previous model, you might notice one glaring problem, and that is, our predictions grow infinitely. As there is a fixed maximum population in California, this would not be possible. Realistically, we would expect the virus to stop spreading once a solution to the spread has been achieved (i.e. A vaccine).

So after noticing this, we also decided to use a polynomial trend model to compare the results and build a more realistic prediction.

Firstly, looking at the cases, we made a sixth-order polynomial model fit our lag one differences data best and plotted this next to the residuals, as shown in Fig 7.

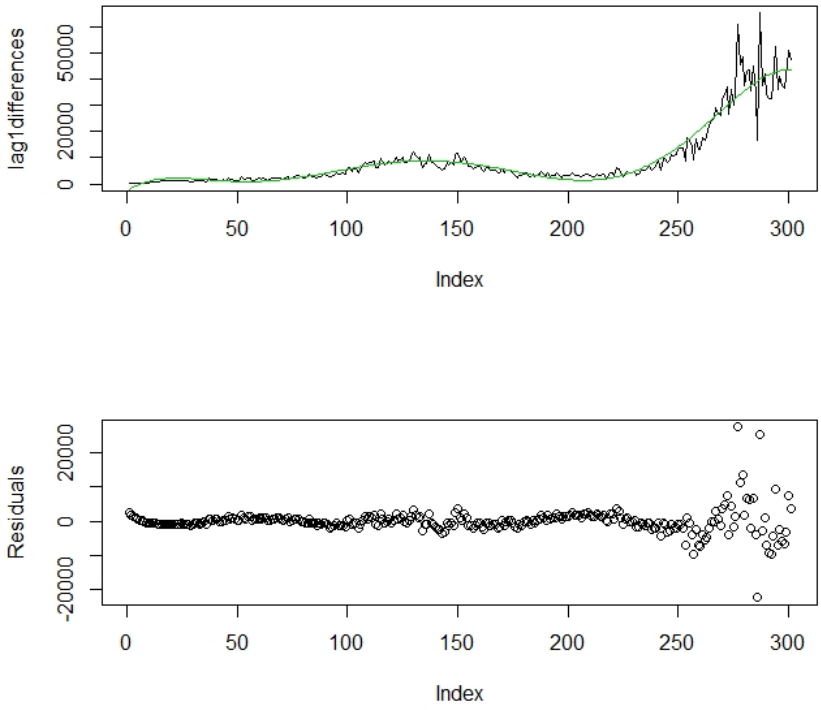

Figure 7: quadratic trend

As we have not yet achieved stationarity, we take the differences between the residuals and preform the ADF test to confirm we now have stationarity. We now proceed with the standard ARMA modelling.

We managed to find a good fit for the residuals from trial and error and like before we found evidence of a seasonal trend. By summing our polynomial (with a constraint: minimum value equal to zero) and ARIMA prediction, we could get our total prediction and plot this, as shown in Fig 8. (The green line is the prediction)

After comparing, we have concluded that a model like this would be more accurate for a long term prediction.

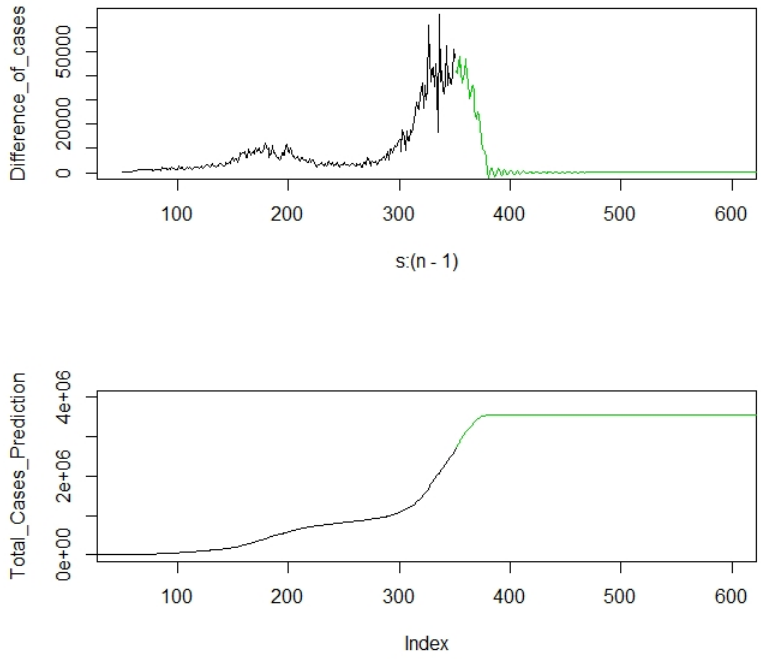

Figure 8: Predictions

4 Comparing models across regions

Following our investigation of California's data, we were interested in finding out if our models applied to other states in the U.S. Before we could start this, we had to split up the states into regions. For the South-West we had: Arizona, New Mexico, California, Colorado, Nevada, Oklahoma, Texas and Utah. And for the North-East we had: New York, Philadelphia, Washington, Boston, Baltimore and Pittsburgh.

We then grouped the dates and summed the cases and deaths in those groups together. The correlation between this data is shown in Fig 9.

From this point, we repeated the methods used above. When comparing models, we noticed consistently higher coefficients in the North-East. This is possibly due to the faster increase of cases in the North-East, as seen in Fig 9.

5 Discussion

The above results indicates the COVID cases and deaths in California will keep growing in next three months but there will be a chance that the exponential growing rate will declined. Also, in the future, the past models of virus will recorded more information about how will the virus spread and cases/deaths increases or decreases. And the past models may be used to measure the stage and the severity of virus in the future.
Forecasting of SARIMA model will be less accurate if the different starting date was been used.
The polynomial trend would have a more reasonable forecast in long-term because of the speciality of the data, and different kinds of data will perform different , like the stock market data, the economic does grow but the short term fluctuation will definitely disappear in the future.