Time Series Analysis of COVID-19 Confirmed Cases in Two Different Provinces of Different Countries

r13250004 邵子軒

December 10, 2024

1 Introduction

COVID-19 first emerged in Wuhan, China, at the end of 2019 and subsequently spread worldwide. We aim to focus on the provinces Shanghai and New South Wales(NSW) in Australia corresponding to the most populous cities in two different countries and conduct a time series analysis.

Using Autoregressive Integrated Moving Average(ARIMA) and Seasonal Autoregressive Integrated Moving Av erage(SARIMA) models, we analyze the time series of COVID-19 confirmed cases. Additionally, we employ spectral analysis to identify potential transmission cycles in the two provinces and compare whether the transmission cycles are influenced by differences in public health policies between the two regions.

2 Data Description

The dataset: Time Series Data Covid-19 Global
https://www.kaggle.com/datasets/baguspurnama/covid-confirmed-global

There are three different types of data from the dataset. We use the confirmed data. The data collected the daily cumulative confirmed cases of the different country provinces from January 22, 2020, to April 24, 2021, obtained from the Johns Hopkins Coronavirus Resource Center.

First, we aim to select a six-month time period for analysis, avoiding periods with rapid increases or decreases in cumulative confirmed cases. Such scenarios often lack relatively stable cycles for prediction. Therefore, we focus on analyzing phases with steady growth. Based on the time periods indicated in **Figures 1 and 2**, we proceed with the analysis.

Next, we calculate the daily confirmed cases. Since the data is cumulative, we use the diff() function in **R** to compute the daily confirmed numbers. The data for Shanghai covers the period from July 5, 2020, to January 1, 2021, while the data for New South Wales spans from October 5, 2020, to April 3, 2021.

After calculating the daily confirmed cases, we plot the data to visualize the distribution.

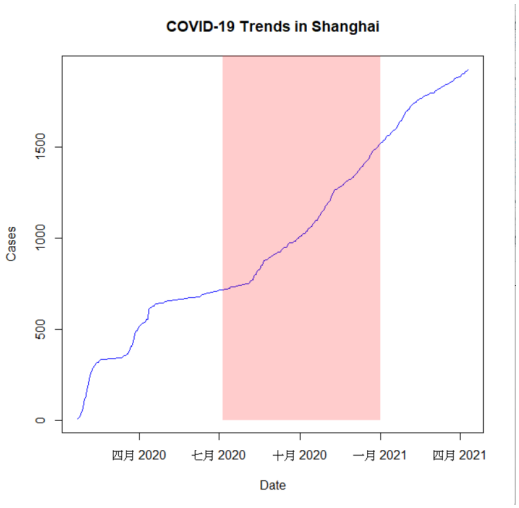

Figure 1: Shanghai

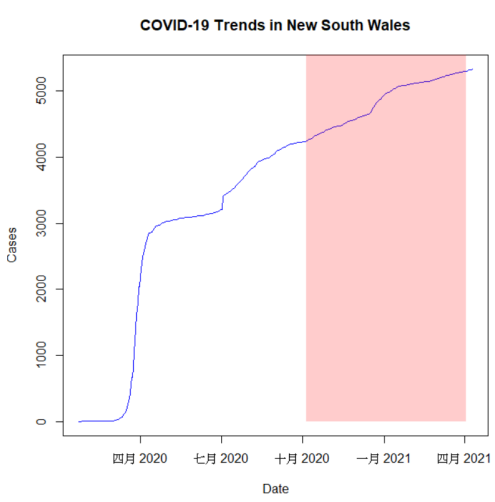

Figure 2: New South Wales

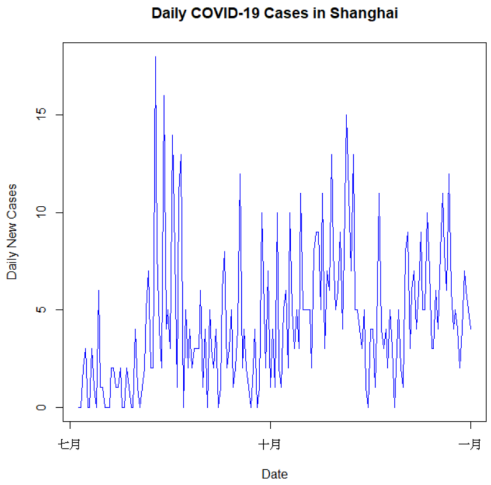

Figure 3: Shanghai

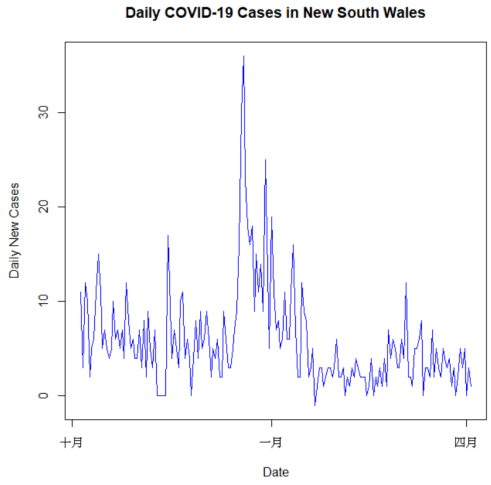

Figure 4: New South Wales

3 Statistical analysis

After extracting the data for analysis, we proceed to the analytical phase. First, from the ACF (Auto Covariance Function) plots of the two datasets, it can be observed that the decay in Shanghai's ACF is less pronounced compared to New South Wales. However, the ACFs of both datasets indicate non-stationarity. To address this, we apply a first-order differencing to make the ACFs relatively stationary.

Next, we examine the ACF and PACF (Partial Auto Covariance Function) to determine the possible parameters for the ARIMA model. From Figures 5 and 6, it is evident that the PACF for Shanghai shows more distinct AR characteristics. Additionally, the ARIMA(3,1,1) model provides better AIC and BIC values than ARIMA(3,0,1). Therefore, we fit the model for Shanghai using ARIMA(3,1,1). Similarly, for New South Wales, the ARIMA(1,1,1) model outperforms

ARIMA(2,1,1) in terms of AIC and BIC values, so we fit the model for New South Wales using ARIMA(1,1,1).

Next, to find out whether there are any significant periodograms and whether these periods can help us use ARIMA to fit the model, we proceed with Spectral Analysis to identify the periodograms of the data. We use the `spec.pgram()` function in **R** with `taper = 0` and `span = NULL`, which applies Fast Fourier Transformations to estimate the periodograms.

The following table shows the top 3 highest periodograms for Shanghai and New South Wales (NSW), respectively.

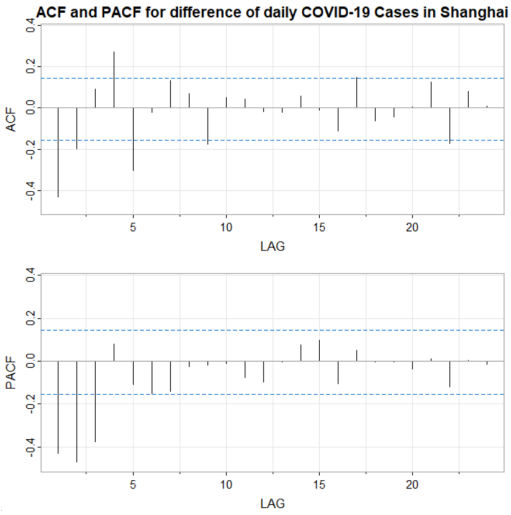

Figure 5: Shanghai

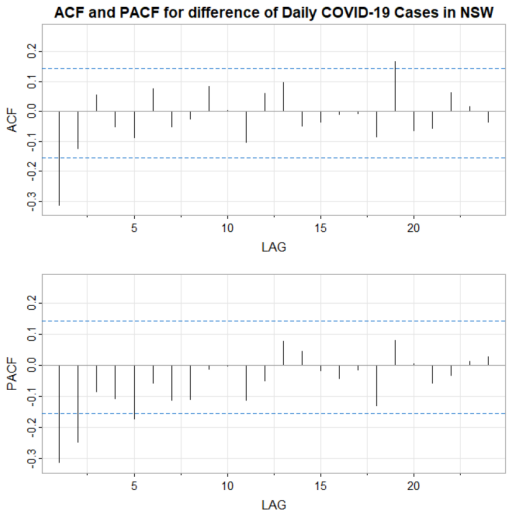

Figure 6: New South Wales

Period	**Shanghai**	**New South Wales**
Period 1	64 days (Frequency: 0.0156)	96 days (Frequency: 0.0104)
Period 2	38.4 days (Frequency: 0.026)	64 days (Frequency: 0.0156)
Period 3	3.49 days (Frequency: 0.2865)	192 days (Frequency: 0.0052)

Table 1: Period and Frequency Comparison for Shanghai and New South Wales

From Table 1, apart from the 3.56-day period, the other periods are too large. Given a sample size of 6 months, approximately 180 days, using a period longer than a month may fail to capture the characteristics of data like COVID-19 case numbers, which exhibit rapid changes and fluctuations.

Based on the ARIMA models discussed, we choose ARIMA$(3,1,1) \times (1,0,1)_3$ for Shanghai's data and ARIMA$(1,1,1) \times (1,0,1)_4$ for NSW's data to fit the models, respectively.Because ARIMA$(3,1,1) \times (1,0,1)_3$ for Shanghai has better AIC and BIC and p-values for Ljung-Box statistic than ARIMA$(3,1,1) \times (1,0,1)_4$.The following **Fig.7,8** are the goodness of fit.

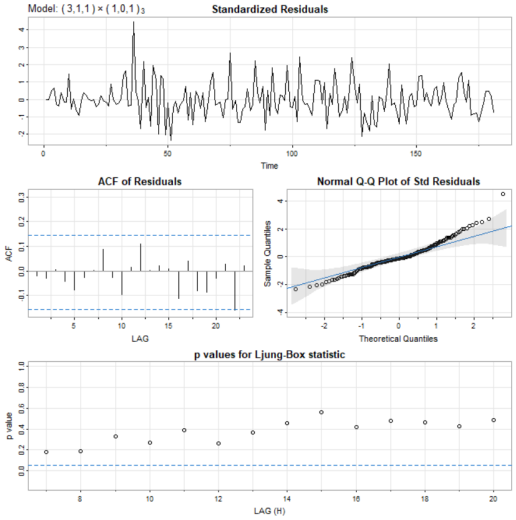

Figure 7: Shanghai

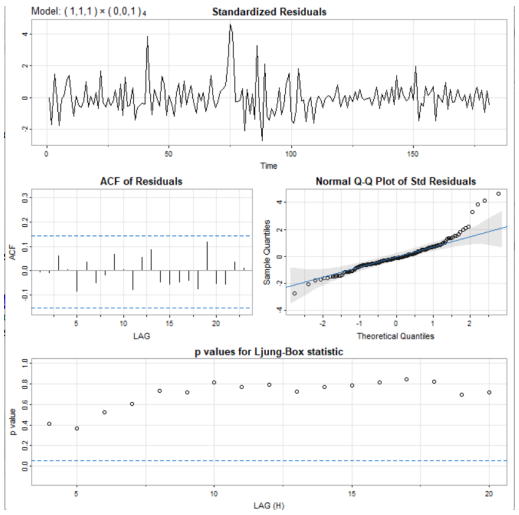

Figure 8: New South Wales

4 Further investigations and interpret the obtained model

First, I would like to interpret the results obtained from the Spectral Analysis and explain why there is such a large difference in the detected periods between the two provinces. Whether these periods are artificially induced or the result of the natural transmission of the virus is an important question. During the selected data period for Shanghai, China's pandemic prevention policies were extremely strict. At that time, most cases were imported from abroad. Due to the stringent lockdown measures and strict testing policies within China, the periods identified by Spectral Analysis are likely reflective of the transmission cycle from abroad to China or the periodic nature of virus testing in China, rather than the domestic transmission cycle within the country.

On the other hand, for New South Wales, the detected periods are mostly multiples of a month. This indicates the potential presence of several transmission cycles influencing the confirmed cases in New South Wales. Unlike Shanghai, Australia's control measures were not as strict, so these periods are more likely to reflect the natural transmission cycles of the virus within Australia, combined with cycles introduced by imported cases.

From **Fig.7,8**, the residual distributions, ACF plots, Normal Q-Q plots, and p-values for the Ljung-Box statistic all align well with our model assumptions. This indicates that the estimated models are relatively reasonable. Specifically for Shanghai, the estimated model is ARIMA$(3, 1, 1)$, meaning that after taking the first-order difference of the data, the current value is influenced by the three preceding time points.

Finally, an intriguing observation is why the 3.56-day period estimated for Shanghai can also fit the data from New South Wales. This might be because the 3.56-day period could reflect the cycle of imported cases. In Australia, there are also patients arriving from various parts of the world. In such a scenario, it is

reasonable that using a 3.56-day or 4-day period could provide a good fit for the data, as both regions may share similar patterns in imported cases.

5 Summary

In this report, we proposed a method combining ARIMA models with Spectral Analysis to identify the transmission cycles of COVID-19 in provinces from different countries. In the first phase, we successfully constructed ARIMA models without including seasonal effects. In the second phase, we integrated Spectral Analysis to identify the corresponding periods for Shanghai and New South Wales, respectively. By leveraging the short cycle detected for Shanghai (3.56 days), we successfully constructed ARIMA models with short seasonal effects.

In the context of COVID-19 transmission, the differences in public health policies across countries are indeed reflected in the transmission cycles. However, the transmission cycle of the virus itself is also evident in the provinces of both countries.