STA457 Final Assignment

Qilong Zeng-1004716943

2022/4/17

Abstract

Today we are in the third year of COVID-19, many people were killed by the pandemic. That bears a question. Will the death rate of COVID-19 decrease in the future so that we can return to normal lives? This report analyses the similar historical events of Hong Kong flu which also caused a sudden increase in the death rate of pneumonia and influenza as severe cases of COVID-19 are often accompanied by pneumonia. We build a SARIMA model on this data to forecast the future periods of the death rate after the shock of the Hong Kong flu by assuming the time series is a stationary process. In addition, we apply Spectral Analysis to illustrate the dominant frequency of the data. Our results show that the negative shock of the Hong Kong flu on the death rate tends to weaken over time. Also, the monthly death rate of pneumonia and influenza has peaked in the yearly cycle; and the rate does not depend on the rate of the previous month. We conclude that the death rate of pneumonia and influenza will return to normal after the shock of COVID-19.

Keywords: Seasonal trend, Time series, Spectral Analysis, COVID-19, Astsa, Pendemic, Statistical Analysis

Intorduction

This data collects monthly pneumonia and influenza deaths per 10,000 people in the US from 1968 to 1978. During the COVID-19 period, many people were killed by COVID-19 pneumonia. Moreover, there are about 15% of cases are severe[1] and it is hard to know when the pandemic will end. Thus, we consider the analysis of monthly pneumonia and influenza deaths in the U.S from 1968 to 1978. The data has 132 observations and is collected by the *Astsa* library [2]. By researching what happened in 1968, we found that there was a flu pandemic called Hong Kong flue which remained as the seasonal flu later [3]. We can see the similarity between COVID-19 and Hong Kong flu. The purpose of this report is to generate some insight into today's

situation from similar historical events. Overall, the impact of the shock by COVID-19 on the death rate of pneumonia and influenza will be Slowly reduced over time. It will be more likely to be remaining as seasonal flu in the future.

Statistical Methods

Data analyze

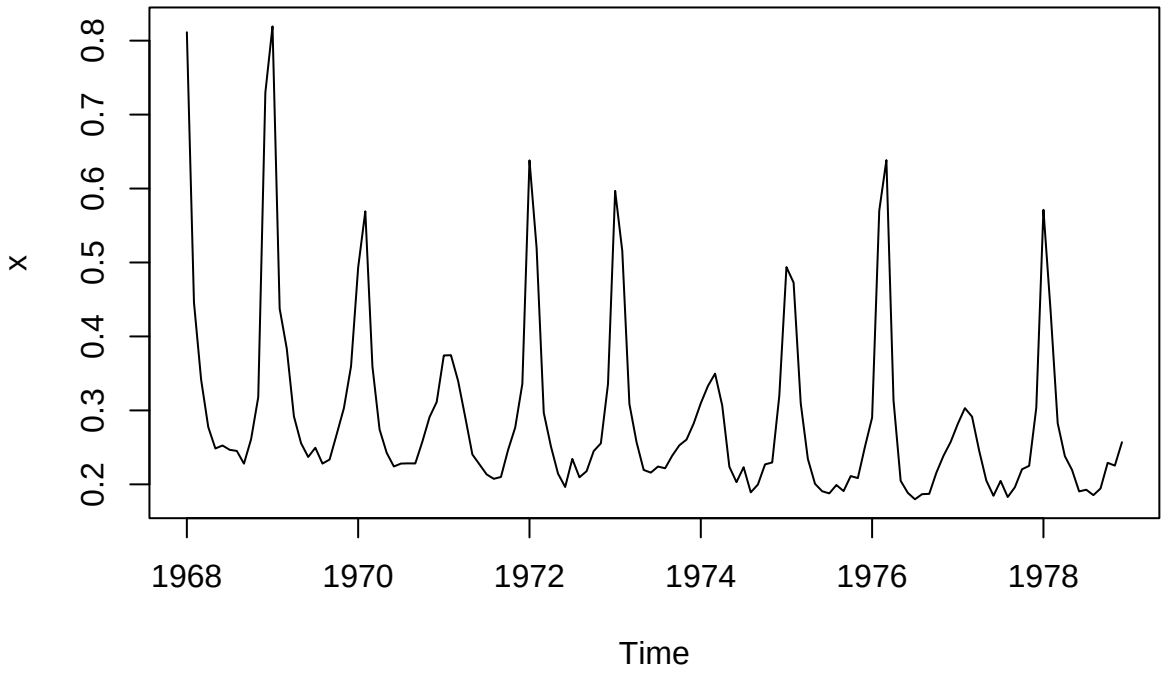


Figure 1: plots of monthly pneumonia and influenza deaths rate over time

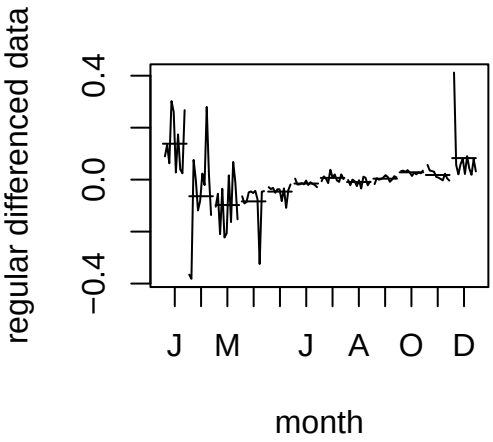


Figure 2: Plots of regular differenced and seasonal differenced data grouped by month

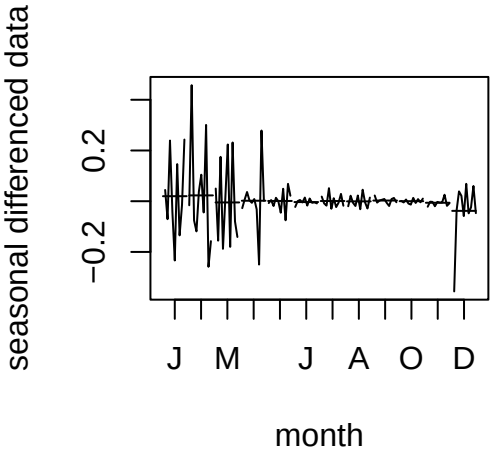


Figure 3: Plots of regular differenced and seasonal differenced data grouped by month

From our original data, we can see the trend and non-constant mean (Figure 1). Therefore, we first differenced data to remove trends and stabilized the mean. However, the differenced data still show trends in seasonal (figure 2). In other words, we expected to see the same variation in the same month of each year recorded. So, we further applied a twelfth-order difference on differenced data. Finally, we constructed the data with a constant mean around zero and without a seasonal trends (figure 3).

ACF/PCAF Analyze

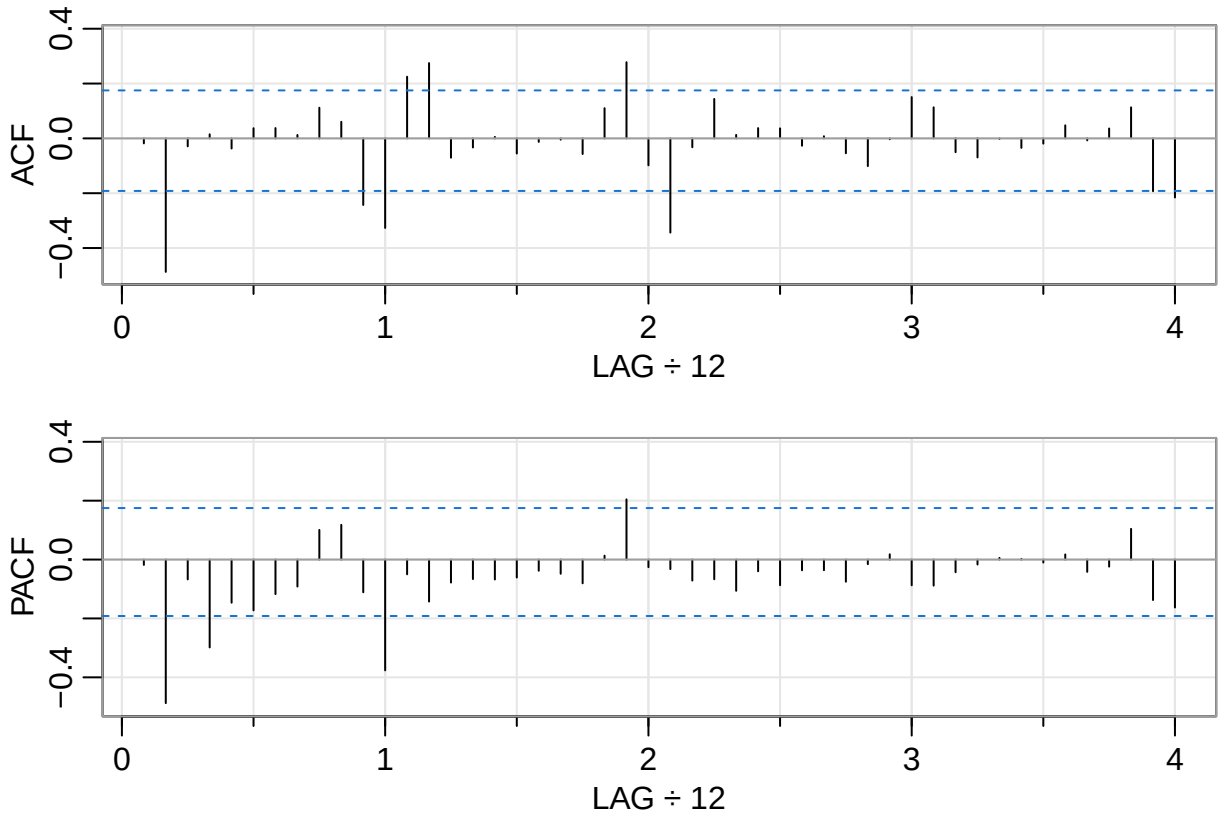


Figure 4: Sample ACF/PACF of regular and seasonal differenced death rate.

Figure 4 is the ACF/PACF of original data after being regular and seasonal differenced $ddx(\nabla_{12}\nabla x_t)$.

From previous steps, we conclude that d=1 and D=1 in our SARIMA model. Next, we are going to propose two models depending on ACF/PACF.

According to the figure, the ACF tails off at lag 1s, 2s, 3s,.... (s=12) and PACF cuts off at lag 1s. Also, the ACF cuts off at lag 2, and the PACF tails off after lag 2. This suggests that $p = 0, q = 2, P = 1, Q = 0$.

It also appears that at the seasons, the ACF cuts off at lag 1s, and PACF tails off. These results imply the

model SMA(1) ($Q = 1, P = 0$).

In conclusion, we propose two model $SARIMA(0,1,2) \times (1,1,0)_{12}$ (model 1)and $SARIMA(0,1,2) \times (0,1,1)_{12}$ (model 2).

Results

Interpretation of Results

Table 1: Estimates of Model 1

	Estimate	SE	t.value	p.value
ma1	-0.2743	0.0756	-3.6279	4e-04
ma2	-0.7257	0.0710	-10.2225	0e+00
sar1	-0.3297	0.0864	-3.8151	2e-04

The estimated $SARIMA(0,1,2) \times (1,1,0)_{12}$ model is

$$\hat{x_t} = \hat{w_t} - 0.2743\hat{w}_{t-1} - 0.7257\hat{w}_{t-2} - 0.3297\hat{x}_{t-12}$$

The results of this model show that the future death rate is negatively related to the noise of the previous two months. Also, it depends negatively on the death rate in the same month last year. From a more practical perspective, the death rate is decreasing from year to year with 0.3297 percent point if all the noise remains the same as last year. Also, all the estimates are statistically significant with a p-value less than 0.05. In other words, we have strong evidence against the coefficient is equal to zero. We can keep all coefficients in this model.

Table 2: Estimates of Model 2

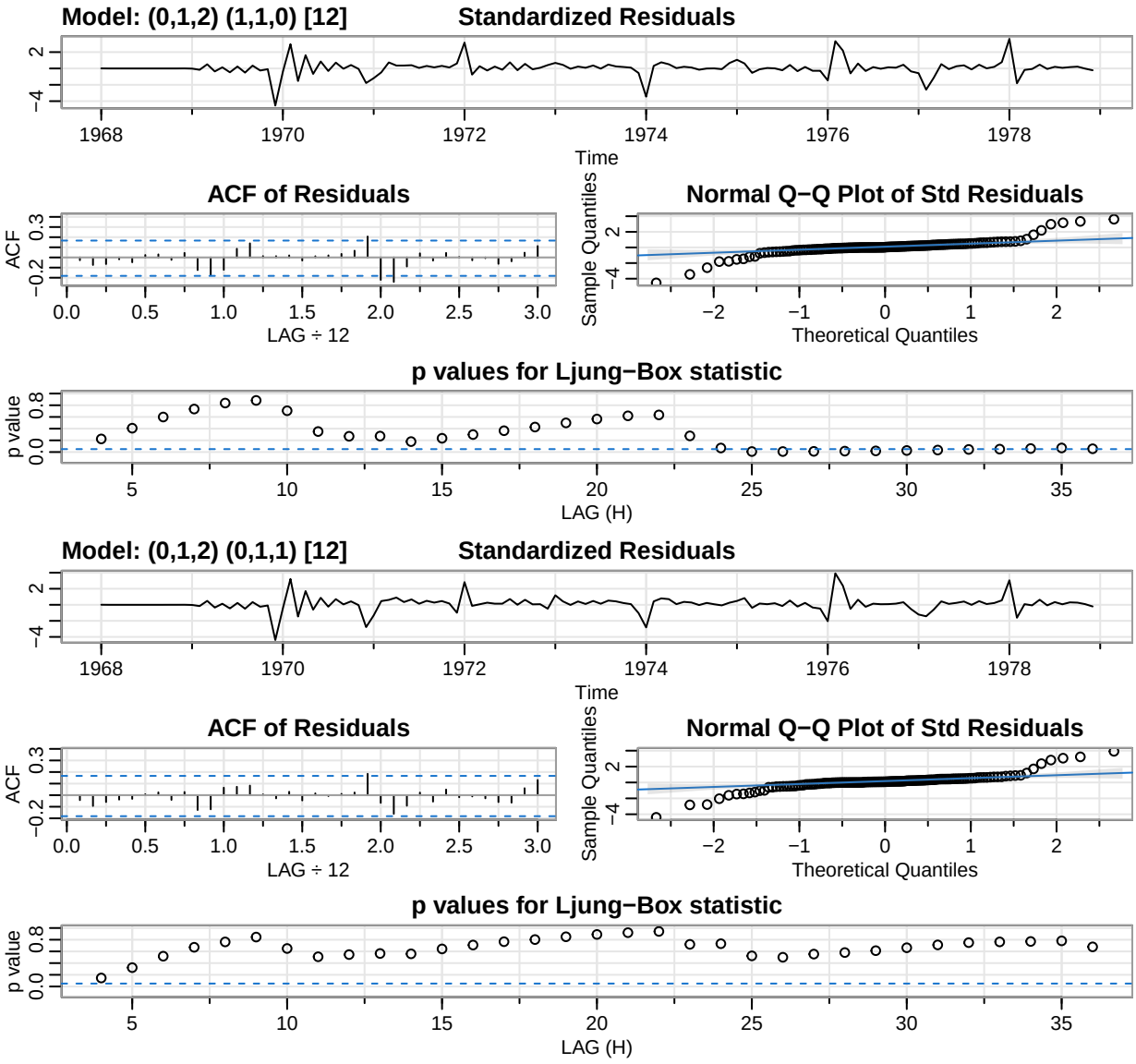
	Estimate	SE	t.value	p.value
ma1	-0.2680	0.1075	-2.4936	0.0141
ma2	-0.7182	0.0960	-7.4790	0.0000
sma1	-0.6427	0.1074	-5.9829	0.0000

The estimated $SARIMA(0,1,2) \times (0,1,1)_{12}$ model is

$$\hat{x_t} = \hat{w}_t - 0.2680\hat{w}_{t-1} - 0.7182\hat{w}_{t-2} - 0.6427\hat{w}_{t-12} + 0.2680 \times 0.6427\hat{w}_{t-13} + 0.7182 \times 0.6427\hat{w}_{t-14}$$

Similarly, the future death rate in this model is also negatively related to the noise of the last two months. In contrast, the death rate further depends on the same month last year and the two months before it. Since those noises are random variables at mean 0, the interpretation is not as practical as the previous model. Moreover, the estimate of MA1 is not statistically significant.

Diagnosis



For both models, there are no obvious patterns from the inspection of the standard residuals. Also, the normal Q-Q plot of Residuals shows a similar trend in both models with several outliers on two tails. However, there are three spikes in ACF of Residuals for model 1 while there is only one for model 2. More importantly, for model 1, almost 1/3 of the p-values for Ljung-Box statistics are at or below the significant level. This suggests that some residuals are dependent. The issue is not existing in model 2. Overall, model 2 fits our data better than model 1.

Model Selection

Next, we compare AIC/AICc/BIC to furthec check which model is the most appropriate model.

Table 3: AIC/AICc/BIC of proposed models

Model	AIC	AICc	BIC
$SARIMA(0,1,2) \times (1,1,0)_{12}$	-2.289264	-2.28751	-2.195848
$SARIMA(0,1,2) \times (0,1,1)_{12}$	-2.38975	-2.387996	-2.296334

AIC/AICc/BIC are all smaller in model 2 than model 1. This suggests that $SARIMA(0,1,2) \times (0,1,1)_{12}$ is better.

Additionally, since ma1 is not statistically significant in $SARIMA(0,1,2) \times (0,1,1)_{12}$. We try two more models which are $SARIMA(0,1,1) \times (0,1,1)_{12}$ (model 3) and $SARIMA(1,1,2) \times (0,1,1)_{12}$ (model 4).

According to the diagnoses, the residuals don't seem independent according to the Ljung-Box statistics because all p-values are below the significance level. At the same time, the diagnosis of model 4 is similar to model 2. Therefore, model 4 is better than model 3.

Table 4: AIC/AICc/BIC of Model 2 and Model 4

Model	AIC	AICc	BIC
$SARIMA(0,1,2) \times (0,1,1)_{12}$	-2.38975	-2.387996	-2.296334
$SARIMA(1,1,2) \times (0,1,1)_{12}$	-2.376624	-2.373675	-2.259854

Model 4 also has one insignificant parameter (ar1). Therefore, model 4 is not strictly better than model 2. Despite that, the AIC/AICc/BIC of model 2 is smaller than model 4. In conclusion, model 2 ($SARIMA(0,1,2) \times (0,1,1)_{12}$) fit our data better than all other models.

Forecast

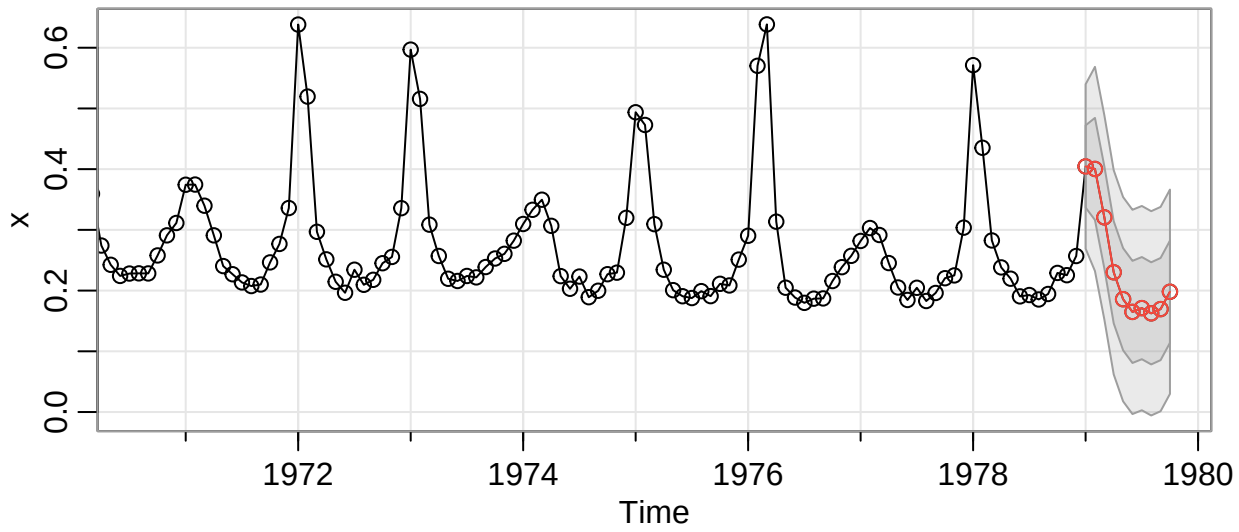


Figure 5: Plots Predicted value and its 95% prediction intervals in next 10-time periods from model 2

Table 5: Predicted value and its 95% prediction intervals in next 10-time periods from model 2

	Predicted value	Upper bound	Lower bound
1	0.4044539	0.5370664	2.718415e-01
2	0.4003425	0.5651458	2.355393e-01
3	0.3206435	0.4854568	1.558301e-01
4	0.2301786	0.3950022	6.535507e-02
5	0.1858217	0.3506554	2.098801e-02
6	0.1648920	0.3297359	4.810772e-05
7	0.1711392	0.3359932	6.285156e-03
8	0.1625118	0.3273760	-2.352441e-03
9	0.1695532	0.3344276	4.678871e-03
10	0.1980936	0.3629781	3.320909e-02

The results of forecasting is shown above. From figure 5, the death rate will decrease from the pick in the first 5 months and increase in the second half of the forecast period. It is pleasurable because Spring is the flu season. With more people get caught by flu, the death rate tend to be increasing. Moreover, the lowest

death rate within the year is the lower than previous rate.

Spectral Analysis

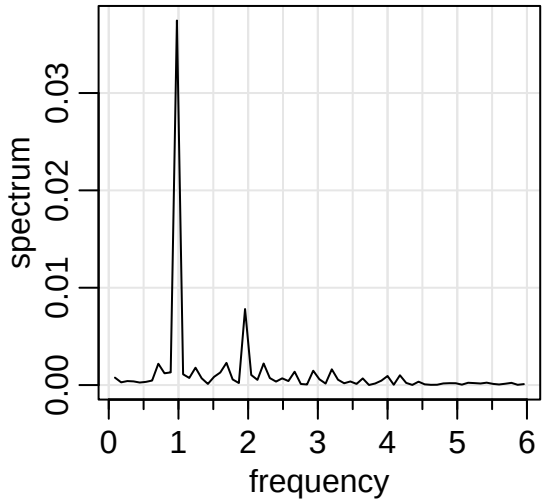


Figure 6: Spectrum density of each frequency

Table 6: First three perdominant periods

frequency	period	spectrum
0.9778	1.0227	0.0374
1.9556	0.5114	0.0078
1.6889	0.5921	0.0023

Table 7: Confidence intervals for the identified periods

Dominant.Freq	Spec	Lower	Upper
0.9778	0.0374	0.0101386	1.4772211
1.9556	0.0078	0.0021145	0.3080835
1.6889	0.0023	0.0006235	0.0908451

According to the spectral summary in figure 7, there is a very clear peak at a frequency of 1. In other words, we notice a narrow-band peak in the yearly cycle. Also, the second dominant frequency is around the

frequency of 2. There is a lower peak at a higher frequency which is a half-year cycle. However, we cannot establish the significance of the first three peaks because the periodogram ordinate of each frequency lies in the confidence intervals of the other two peaks.

Discussion

Overall, our model shows that the future death rate is negatively related to the noise of the last two months, the same month last year, and two months before it. The forecast suggests that the lowest death rate in the future will be the lowest of all previous years. We conclude that the effects of the immediate increase in death rate trend caused by Hong Kong flu are weakened from year to year. The variation of the death rate tends to depend on seasons. It reaches a peak in the yearly cycle. From the insight of this report, the impact of COVID-19 on pneumonia and influenza death rate will keep decreasing in the future. It is also possible that COVID-19 will become the seasonal flu as Hong Kong flu.

Although, we tried to find the best model to fit our data. The final model we used for the forecast still had limitations. There were several outliers on two tails of the Normal Q-Q Plot of Residuals. It means there were departures from the normality assumption. It is possible that the process of this data is not stationary which means a SARIMA model is not appropriate. Therefore, other mathematical models could be a better fit. Also, we cannot establish the significance of the first three peaks. This means our data does not appear significant periodic pattern.

All analysis for this report was programmed using `R version 4.0.5`[4] with packages *knitr*[5] and *astsa*[2]

Reference

[1] Ratini, M. (2022, January 25). Pneumonia and coronavirus: Does everyone with covid-19 get pneumonia? WebMD. Retrieved April 9, 2022, from https://www.webmd.com/lung/covid-and-pneumonia#1

[2] David Stoffer (2021). astsa: Applied Statistical Time Series Analysis. R package version 1.14. https://CRAN.R-project.org/package=astsa

[3] Wikimedia Foundation. (2022, March 8). Hong Kong flu. Wikipedia. Retrieved April 9, 2022, from https://en.wikipedia.org/wiki/Hong_Kong_flu

[4] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[5]Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.36.