A Time Series Analysis on Monthly Pneumonia and Influenza Deaths in the US

Winter 2021 Sta457 Final Project

Author: Chaerin Song

Abstract

This analysis aims to identify the best fit SARIMA model for the monthly pneumonia and influenza death in the United States from 1968 to 1978 and to forecast the 12 months death rate in the year of 1979. After transforming the data and selecting couple most probable time series models by investigating plots, autocorrelation and partial autocorrelation functions, a seasonal ARIMA model with parameters (2, 0, 0) x $(0, 1, 1)_{12}$ was selected based on the diagnostic plots and p-values of the estimated coefficients. The model produced reasonable forecast for the next 12 months, and the periodogram analysis identified $\frac{1}{12}$ as the most predominant period. Overall, a 12 months (or yearly) cycle was identified in the data with most peaks in January, which supports the general knowledge that people are exposed to flu the most during the winter.

Introduction

After a year into the pandemic, experts claim that COVID-19 "may become feature of lives, like seasonal flu" (*BBC News*). By looking at the seasonal trend in historical flu death rates, I hoped to draw out meaningful analyses on epidemic behavior to better understand the new virus. The "flu" data from the R "astsa" library was used for analysis. The data consists of total 132 data inputs on the number of monthly pneumonia and influenza deaths per 10,000 people in the United States for 11 years, 1968 to 1978. Also, it is widely known that for the most years flu activity peaks between December and February (*CDC*). By conducting a full time series

analysis, I hoped to further support the yearly cycle of flu and to investigate the presence of additional cycle periods.

Statistical Methods

The analysis will follow these steps in order: 1) data transformation and stationarity, 2) model proposal, 3) model selection and diagnostics, 4) forecasting, and 5) spectral analysis. The data transformation includes finding a proper transformation using the shape of the time series plot, autocorrelation plot, and Box-Cox functions in R, and the stationarity can be reached through differencing the series. Then we observed ACF and PACF plots of the processed data to figure out the dependence order of both the seasonal and non-seasonal ARIMA, and also the most proper autoregressive, moving average, and seasonal parameters to build top 3 models. Then, each model will be examined through the diagnostic plots (residual plot, QQ plot, ACF of residuals, and Ljung-Box statistics) and estimated parameters with their p-values to select the best fit model (SARIMA $(2, 0, 0) \times (0, 1, 1)_{12}$). With this model, we forecast the flu mortality rates of the future 12 months and their prediction intervals. Finally, we conduct a spectral analysis to figure out the most predominant periods in the original flu series.

Results

- **Data Transformation and Stationarity**

  The unprocessed flu data is clearly seasonal and non-stationary, as shown in Figure 1.

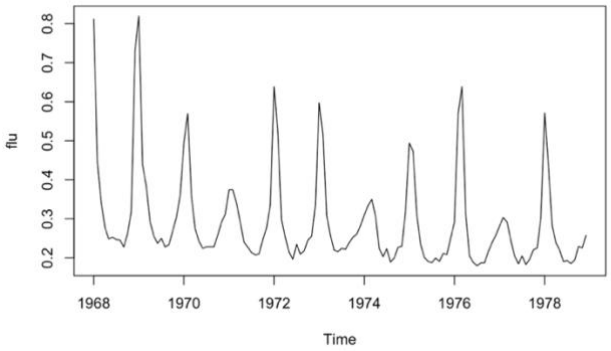


Fig. 1. *Plot of the unprocessed flu data.*

The R code `BoxCox.lambda(flu)` returns a lambda value very close to -1, which indicates the necessity of inverse transformation ($\frac{1}{flu}$).

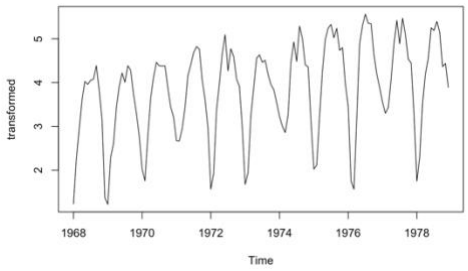


Fig. 2. 1. *Plot of* $\frac{1}{flu}$

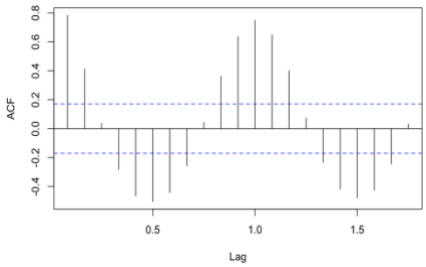


Fig 2. 2. *acf of* $\frac{1}{flu}$

Looking at Figure 2.1, we notice that the variance is not quite stable throughout the data. As the variance seems to be slightly increasing over time, I used Box-Cox class of power transformations with $\lambda = 0.25$ on the inversely transformed flu data.

Now that we have suitably transformed the data and stabilized the variance, we can apply differencing techniques to make the data look stational. In Figure 2.1, it is notable that the minimum peaks occur yearly, or for every 12 data inputs. Figure 2.2 is the corresponding autocorrelation function of the transformed data, and the systematic oscillating pattern indicates the necessity of seasonal differencing. First, I differenced the inversely

transformed data by 12 to normalize the seasonal pattern. Figure 3.1 is the plot of the seasonally differenced data, and Figure 3.2 is the corresponding autocorrelation function to indicate if any further differencing is necessary.

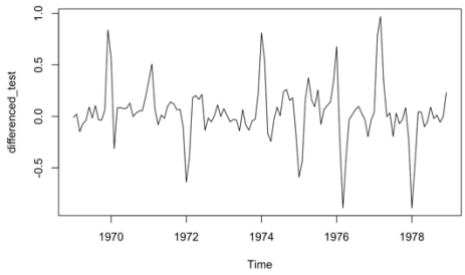


Fig. 3.1. *Plot of the differenced data*

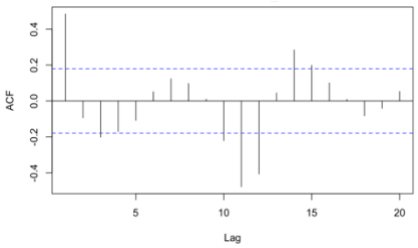


Fig. 3.2. *acf of the differenced data*

If there is a slowly decaying trend in the ACF plot, we should consider a non-seasonal differencing. In the Figure 3.2, we do not observe a slow decay of the sample ACF. Hence, I stopped transforming the data at the seasonal differencing of 1 and no non-seasonal differencing. As the mean and the variance seem to be constant with no more differencing needed, we now can assume that the processed data is stationary and ready to fit a SARIMA model.

- **Model Proposal**

  In order to propose possible models, it is important to review the ACF and PACF plots of the processed data. In Figure 4 below, the non-seasonal ACF tails off, and seasonal ACF could be either interpreted to be 1) tailing off or to 2) spike at 11 and 12 (about lag 1) and cuts off (top figure). Meanwhile, the non-seasonal PACF cuts off after 2, and the seasonal PACF tails off (bottom figure). Since the seasonal ACF tails off or cuts off at lag 1, and the seasonal PACF tails off, it is suggested to use seasonal ARMA or MA model with the seasonal difference at 12. And since the non-seasonal ACF tails off and the PACF cuts off, it is our primary option to use a non-seasonal AR model with the difference of 0.

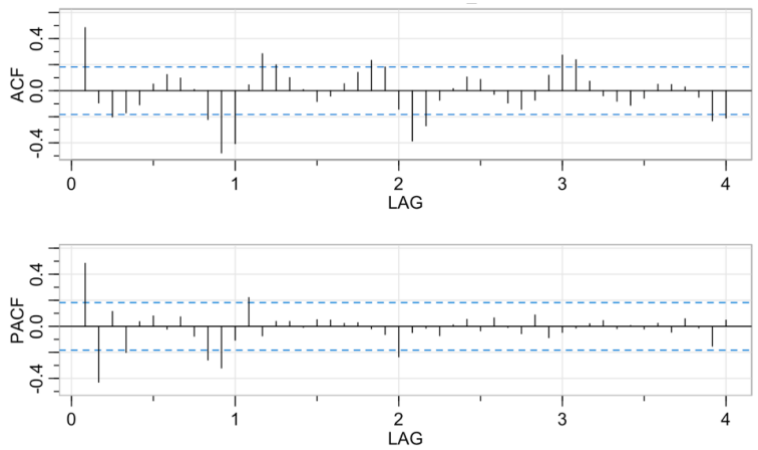


Fig. 4. *ACF and PACF plots of the differenced flu data*

With this information, here are the top 3 proposed models with their estimated parameters, the significance of the estimates, and the contextual interpretations:

1) **SARIMA (2, 0, 0) x (1, 1, 1)$_{12}$**

	Estimate	SE	t.value	p.value
ar1	0.6541	0.0872	7.4968	0.0000
ar2	-0.4250	0.0870	-4.8837	0.0000
sar1	0.0784	0.1318	0.5954	0.5528
sma1	-0.9128	0.2248	-4.0599	0.0001
constant	0.0034	0.0005	6.2972	0.0000

Here, every parameter but the seasonal AR1 is significant. Flu death rate of a month relies on the two previous months' death rates, along with the year prior death rates and the errors as well.

2) **SARIMA (2, 0, 0) x (0, 1, 1)$_{12}$**

	Estimate	SE	t.value	p.value
ar1	0.6434	0.0860	7.4841	0
ar2	-0.4176	0.0867	-4.8169	0
sma1	-0.8494	0.1234	-6.8843	0
constant	0.0034	0.0005	6.2724	0

For this model, every parameter estimate is significant. In a scientific context, it suggests that the flu death rate of a month relies on the two previous months' death rates, along with the year prior prediction errors as well.

3) **SARIMA (2, 0, 0) x (0, 1, 2)$_{12}$**

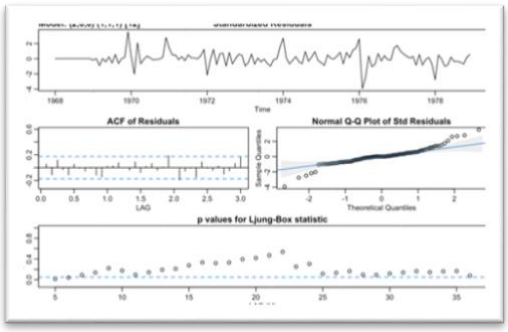
	Estimate	SE	t.value	p.value
ar1	0.6576	0.0871	7.5502	0.0000
ar2	-0.4275	0.0869	-4.9200	0.0000
sma1	-0.8322	0.2584	-3.2207	0.0017
sma2	-0.1013	0.1556	-0.6509	0.5164
constant	0.0034	0.0005	6.2953	0.0000

This time, every parameter estimate is significant except for the seasonal MA2 parameter. The model suggests that the death rate of a month relies on the two previous months' death rates like other models, along with the year prior and two years prior prediction errors.

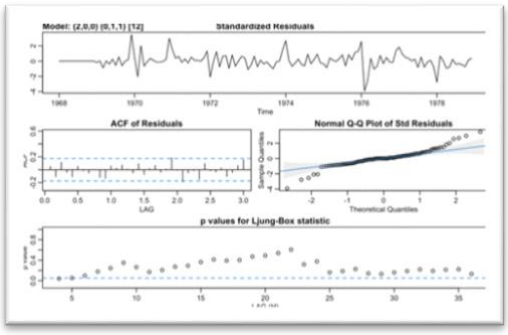
- **Model Diagnostics and Selection**

Here are the diagnostic plots of the 3 proposed models and AICc and BIC table:

SARIMA (2, 0, 0) x (1, 1, 1)$_{12}$:



SARIMA (2, 0, 0) x (0, 1, 1)$_{12}$



SARIMA (2, 0, 0) x (0, 1, 2)$_{12}$:

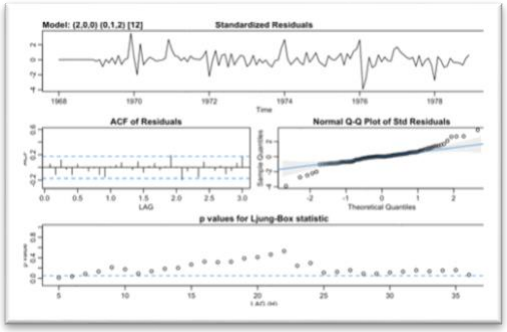


Table 1. *Model Selection Criteria*

Model	AICc	BIC
SARIMA (2, 0, 0) x (1, 1, 1)$_{12}$	- 0.4348	- 0.3108
SARIMA (2, 0, 0) x (0, 1, 1)$_{12}$	- 0.4485	- 0.3446
SARIMA (2, 0, 0) x (0, 1, 2)$_{12}$	- 0.4360	- 0.3120

For all three models, the standardized residuals plots do not have a trend or notably inconsistent variances. Although the extreme values divert away from the normal distribution, most points are lying within the normal range. There is barely any significant value of autocorrelation in the ACF plots of residuals, and the p value plots for Ljung-Box statistic suggest that there is almost no autocorrelation left in the residuals. Thus, it is necessary to look at the AICc and BIC values in the model selection criteria (Table 1). The smaller the values, the better model it probably is. Out of three models, the second model SARIMA (2, 0, 0) x $(0, 1, 1)_{12}$ has the smallest AICc and BIC. Therefore, SARIMA model of p = 2, d = 0, q = 0, P = 0, D = 1, Q = 1, and S = 12 is selected as the best fit model for the processed flu data. In addition, all the estimated parameters of the model were tested significant in the previous section, so all parameters are to be included in the final model.

Final SARIMA model:

$$x_t = 0.6434x_{t-1} - 0.4176x_{t-2} + x_{t-12} - 0.6434x_{t-13} + 0.4167x_{t-14} + w_t - 0.8494w_{t-12}$$

(where x_t is the inverse of the number of deaths from flu per 10000 people at time t.)

- **Forecasting**

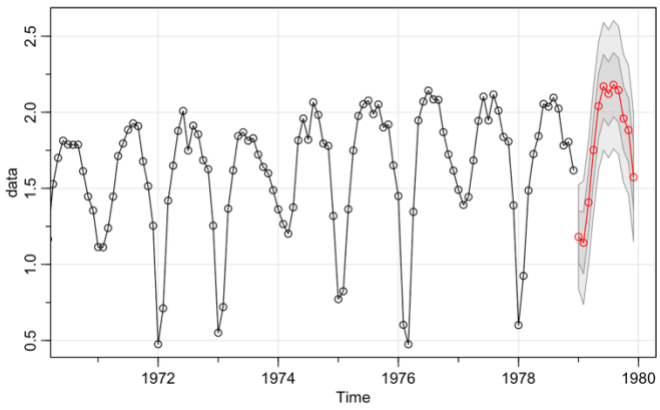


Fig 5. 1. *12 months ahead forecast of the processed data*

	Lower	Upper
Jan 1979	1.268427	1.323976
Feb 1979	1.245952	1.311980
Mar 1979	1.511535	1.577560
Apr 1979	1.857695	1.925410
May 1979	2.148230	2.216621
Jun 1979	2.277530	2.345921
Jul 1979	2.229282	2.297794
Aug 1979	2.288435	2.356995
Sep 1979	2.252189	2.320748
Oct 1979	2.066816	2.135380
Nov 1979	1.990561	2.059128
Dec 1979	1.681134	1.749680

Fig 5. 2. *Prediction Interval*

With the selected model, here is the forecast of the inverse and Box-Cox transformed of the death rate from flu for the future 12 time periods ahead, from January to December 1979. In Figure 5.1, the red line is the predicted values, which looks very consistent with the previous pattern in the data. Figure 5.2 is the list of all the 95% prediction intervals of the 12 future time points, and the length of each interval is impressively small. Given that our model is a good fit to the data, this result is not too surprising.

Now that we saw the accuracy of the predictions, we could reverse the process (inverse and Box-Cox transformations) to the prediction results to better understand the data in context. With this reversion, the prediction is for our original flu data.

1979	Jan	Feb	Mar	Apr	May	Jun
prediction	0.3552	0.3661	0.2993	0.2339	0.1922	0.1766
1979	Jul	Aug	Sep	Oct	Nov	Dec
prediction	0.1823	0.1754	0.1796	0.2030	0.2137	0.2653

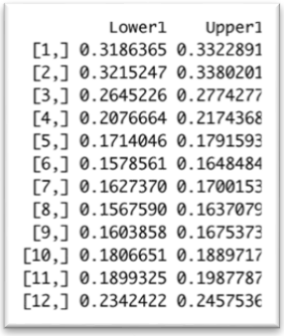
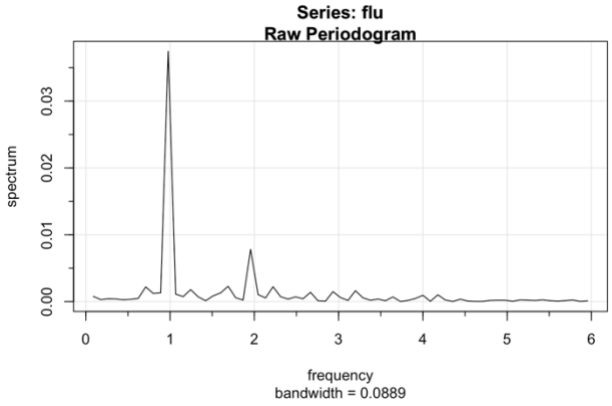


Table 2. *Prediction of the number of deaths from flu per 10,000 people in U.S (1979)*

Figure 6. *95% prediction intervals of the 12 points predictions of original flu data (right)*

Table 2 is the list of prediction results of the original data, and as we expected and observed from the sample data, 3 months of winter from December to March have the highest death rate. And as expected, the 95% prediction intervals for the original flu series predictions are very small in their lengths (between 0.005 and 0.015).

- **Spectral Analysis**



1st Predominant Period	0.08333 (= $\frac{1}{12}$)
2nd Predominant Period	0.16667 (= $\frac{1}{6}$)
3rd Predominant Period	0.007576 (= $\frac{1}{132}$)

Figure 7. *Raw periodogram of the original flu data (left)*

Table 3. *Top 3 predominant periods of the original flu data (total 132 points) (right)*

Figure 7 and Table 3 indicate that top 3 predominant periodic components of the flu data are 12, 6, and 132, respectively from the best to the third best. Note that the third predominant periodic component (132) is meaningless, since the total input count is 132 in this series. The most predominant periodic component is a 12-month cycle, which is what we would expect in a yearly seasonal series. The next predominant periodic component is interestingly a 6-month cycle, although it is hard to distinguish from our plot.

Period	95% Confidence Interval	
	lower	upper
0.08333	0.01015	1.47903
0.16667	0.00212	0.30863

Table 4. *95% confidence intervals for the top 2 identified periods*

In Table 4, the 95% confidence intervals for the frequencies are not that narrow. The lower value 0.010 of the first period indicates that the actual most predominant periodic component of the series is at least more than a month cycle with a 95% chance. Based on the length of this CI, it is hard to conclude that the highest peak in the spectrogram is significant.

Discussion

Our flu mortality model predicts the flu death rate of a month depending on the death rates of two previous months, a year prior, and 13 and 14 months prior, along with on the error of the year ahead rate.

It was very difficult to find anything more specific than a yearly mortality rate from open sources. If the data is available, we could further investigate how accurate our 1979 prediction is. Our main limitation comes from the nature of pneumonia and influenza. Although the mortality rate generally follows a seasonal trend, there are a lot of other factors that act in the mortality rate of a year, such as the vaccine's efficacy, weather, and other epidemics of the year. Another limitation comes from the fact that after differencing by 12 on our processed model, it still does not look perfectly stationary, as if there is a further seasonal dependence to be investigated on. Furthermore, as the yearly vaccinations usually play a huge role in the mortality rates, it is hard to tell what our seasonal model really means, whether it indicates the year's vaccine efficacy, or the mortality of the year's variant. With more virologic understanding, the model could be revied to indicate which year's vaccination would be considered failure, or how the vaccine failure affects the mortality rate.

Works Cited

"Coronavirus Doctor's Diary: Will Covid Be with Us Forever, like Flu?" BBC News, BBC, 13 Feb. 2021, www.bbc.com/news/health-56047489.

"The Flu Season." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 July 2018, www.cdc.gov/flu/about/season/flu-season.htm.