Time Series Analysis on U.S. Citizen Air Travel to Canada

Leanne Lee

Introducion

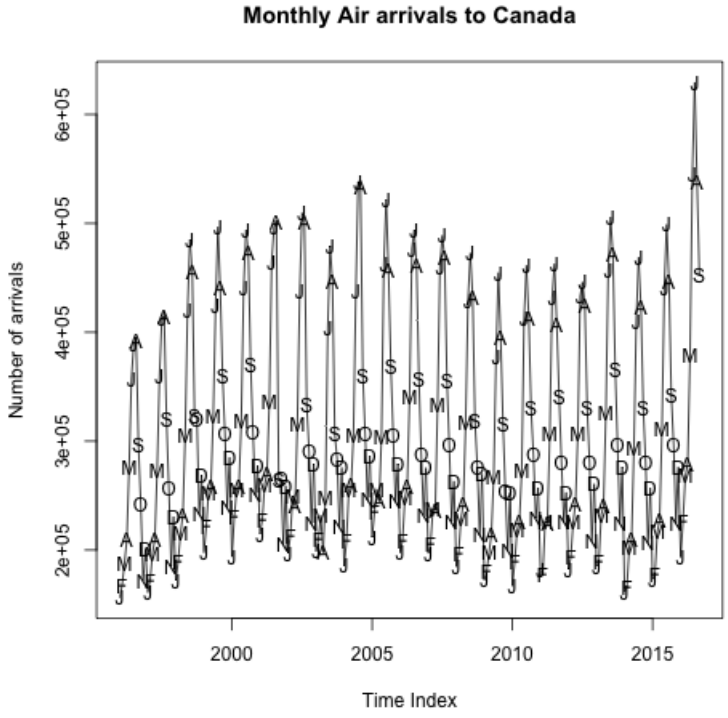
The scientific question motivating my work is whether there exists a seasonal pattern for U.S.citizens to travel to Canada by air. The inspiration of my research question came from my summer traveling experience to Toronto, Canada. I was having a hard time to book bus tours to Montreal and Quebec because of the massive popularity with U.S. tourists during summer. Since Canada is one of the closest countries across the border of the United States, many U.S. citizens consider Canada a great place to spend their holidays. In the West coast, we have Vancouver and Victoria, while the East coast has Toronto, Montreal and Quebec.

Dataset

The dataset comes from the U.S. Department of Commerce, National Travel & Tourism Office. NTTO The dataset contain U.S. outbound travel by world regions from Jan 1996 to Sep 2016. I filtered out the dataset containing only Canada region. There are 249 monthly data points throughout these twenty years.

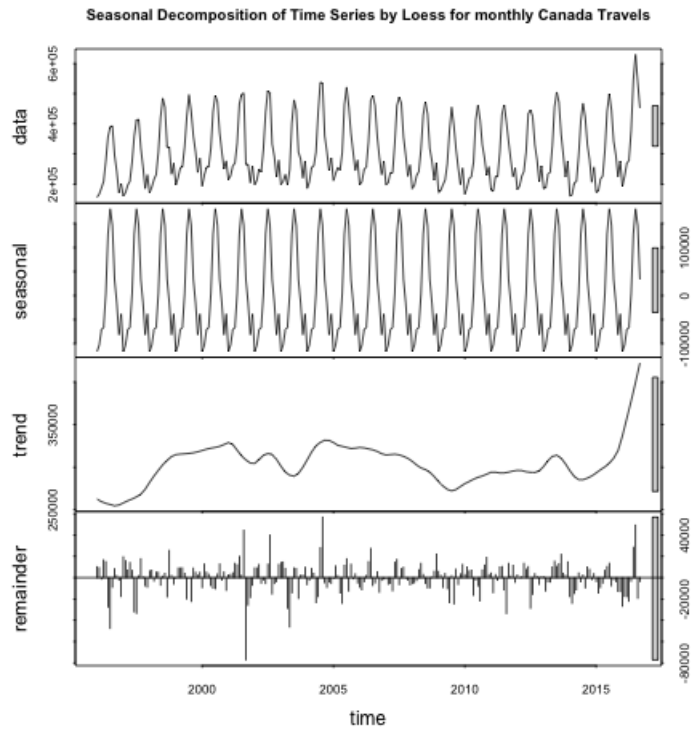
Exploratory Data Analysis

First, I plotted the time series for the monthly air arrivals from the U.S. to Canada. By examining the time series, the plot looks seasonal with high peaks in June, July and August. Students usually travel during their summer break and this is the main drive of the increase in air travel during these months. There is also a smaller peak in between of each cycle. These small peaks happens in December, which is another high season for students to travel to Canada during Christmas. The time series plot does not get effected when the variation increase with the level of the series, so I do not need to take the log of the time series.



After the basic inspection on the time series, I ran the stl() command for a seasonal decomposition of Time Series by loess. I first inspected the data section where the data looks stationary. I considered the bar on the right hand side as one unit of variation. The bar on the seasonal panel is slightly bigger than the data panel, meaning the seasonal effect is larger than the variation in the data. In the trend panel, it has a larger variation box than the data and seasonal panels, meaning the variation attributed to the trend is much smaller than

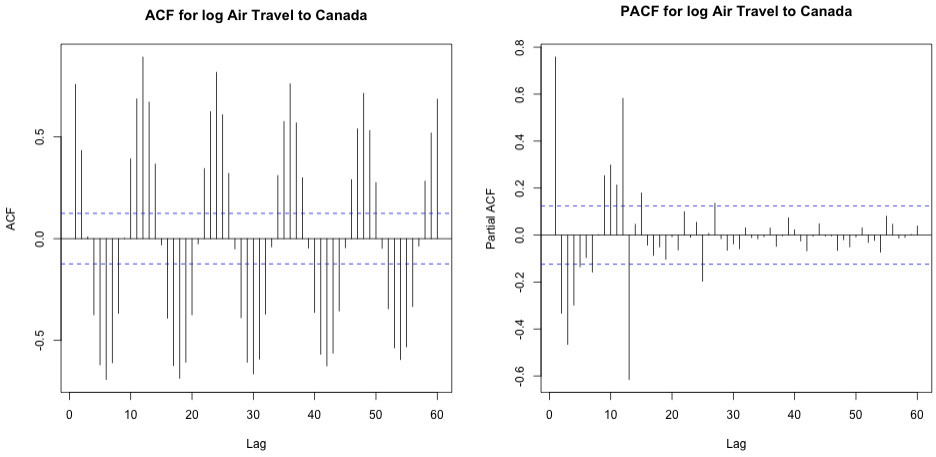
the seasonal component. It also indicated that the trend is not dominating, but there is an upward trend towards the end. The spike upward trend of air traveling to Canada in July 2016 is due to the low Canadian dollar, causing a boost to tourism in Canada. The remainder panel shows there are some high residuals in year of 2001 to 2005 and the year of 2016.



Seasonal ARIMA Model

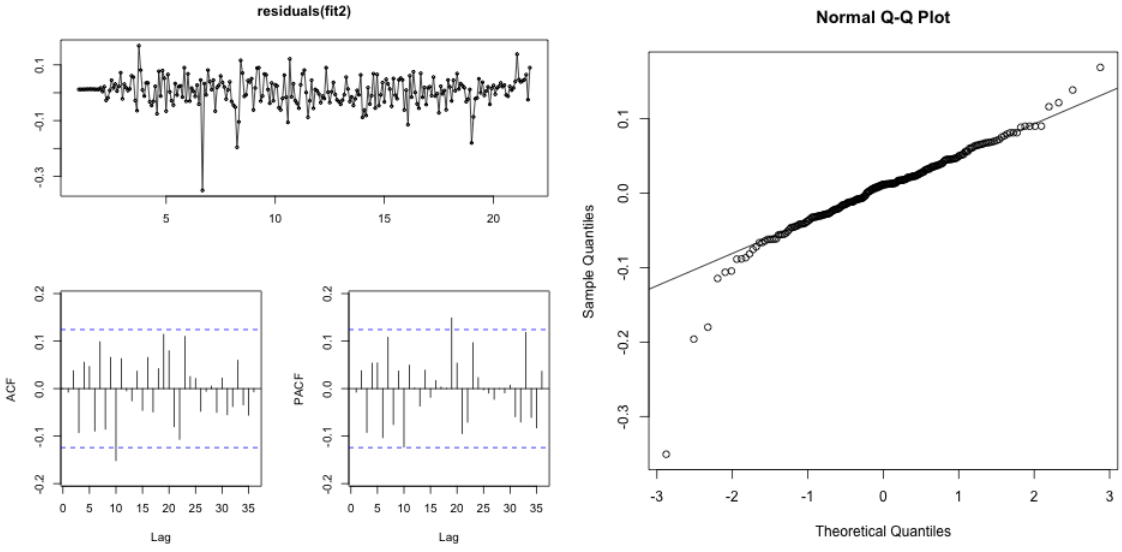
Based on the results on ACF and PACF, I found that there are seasonal autoregressive and moving average effects. So I ran the ARIMA model with changing order of p,d, q and seasonal orders. First, I ran the auto arima to get a rough estimate of my orders in seasonal ARIMA. Since the original series looks stationary, I do not need to consider a high order of differencing (**d**).

ACF and PACF



ARIMA tested models: ARIMA $(2,0,2)X(2,1,2)_{12}$ ARIMA $(2,1,1)X(0,1,1)_{12}$ ARIMA $(2,0,1)X(0,1,2)_{12}$ ARIMA $(1,0,2)X(0,1,2)_{12}$ ARIMA $(1,1,1)X(0,1,2)_{12}$ With the strong and stable seasonal pattern, I used (0,1,2) as the seasonal order. From the list of ARIMA models I tested, the best fitted ARIMA model with the lowest AIC is **ARIMA** $(2,0,1)X(0,1,2)_{12}$

Residuals of the Model



After fitting the model, it is important to diagnostic the residuals of the ARIMA model. The blue line indicates the 95% confidence interval under the null hypothesis for white noise.

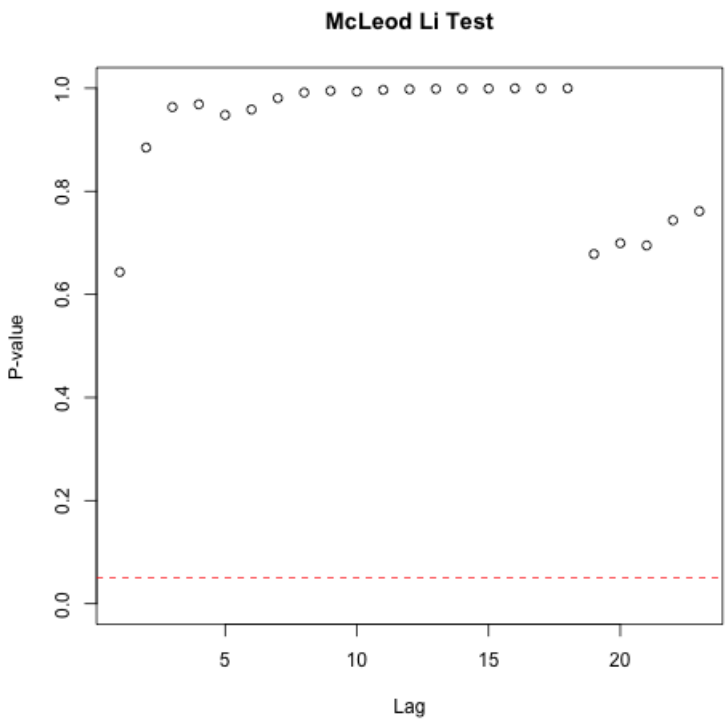
Both ACF and PACF show the majority of residuals are under the blue threshold. so they pass the residual tests. The residuals of fitted models has a few outliers in 2001, 2003 and 2014. I will detect these outliers in the later section. I also examined the QQ plot to check the normality of the residuals. The plot shows there are a few outliers in the beginning. The most significant outlier is the first point, which caused by the extreme high peak during July 2016. I will use outliers detection methods to find out the exact index of the outliers.

Detecting Outliers

```
##             [,1]      [,2]       [,3]
## ind     69.0000 88.000000 217.000000
## lambda1 -7.2771 -4.068404  -3.735382
```

From the residuals plot, I decided to use **detectAO()** and **detectIO()** to determine the outliers. There is no result for detecting additive outliers (AO), so I examined the innovative outliers (IO) and found 3 outlier in index 69, 88 and 217. The biggest magnitude comes from index 69, which is September 2001. The main reason of the decreased in traveling to Canada is because of the 9/11 attack. People are being more skeptical of taking flights during September 2001. The second index of 88 (April 2003) had a drop in air traveling to Canada due to the Severe Acute Respiratory Syndrome (SARS) virus outbreak in Toronto. This life threatening virus raised concerns for U.S. citizens to travel to Canada. Although the detectIO function can detect a few outliers from the dataset, it was not able to detect the last outlier in July 2016. That outlier shows a drastic increase in air traveling to Canada, as well as causing the increasing trend.

McLeod-Li test



I ran the McLeod-Li test to see if the p-values reject my null hypothesis. All points are above 0.5, thus I don't reject the null hypothesis and it is unnecessary to use the arch garch model.

GARCH Model

Although the McLeod-Li test shows there is not a need for arch-garch test, I want to reassure my assumption by checking it with the GARCH model. GARCH model suspect heteroskedasticity over time. From my original time series, it show that there is only one bump at the end of the series. Thus, I used the standard approach for modeling volatility of Garch(1,1) on my residuals of the fitted model.

```
summary(g)
```

```
##
```

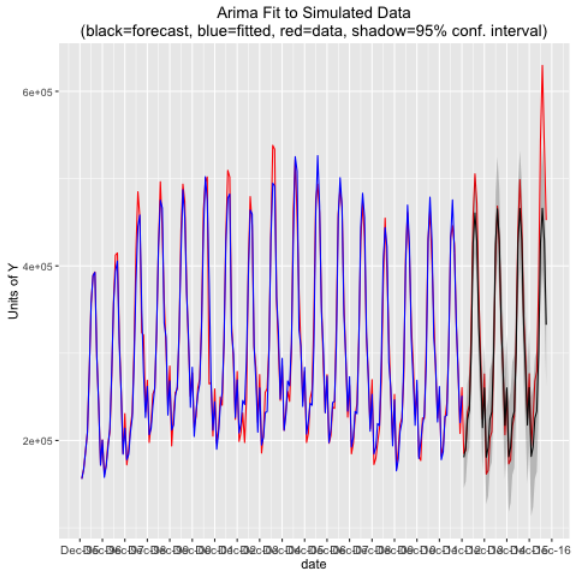
```
## Call:
## garch(x = resModel, order = c(1, 1))
##
## Model:
## GARCH(1,1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7203 -0.4608  0.2161  0.6887  3.0750
##
## Coefficient(s):
##     Estimate  Std. Error  t value Pr(>|t|)
## a0 2.427e-03   7.090e-04    3.423 0.000618 ***
## a1 1.424e-01   1.082e-01    1.316 0.188124
## b1 4.189e-14   2.615e-01    0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Diagnostic Tests:
##  Jarque Bera Test
##
## data:  Residuals
## X-squared = 933.35, df = 2, p-value < 2.2e-16
##
##
##  Box-Ljung test
##
```

```
## data:  Squared.Residuals
## X-squared = 0.017711, df = 1, p-value = 0.8941
```

The estimate of b1 is extremely small. When I applied b1 into the Garch(1,1) formula above, ß1 became insignificant and shows that there is no autoregressive effect. Therefore, GARCH is not useful in this case.

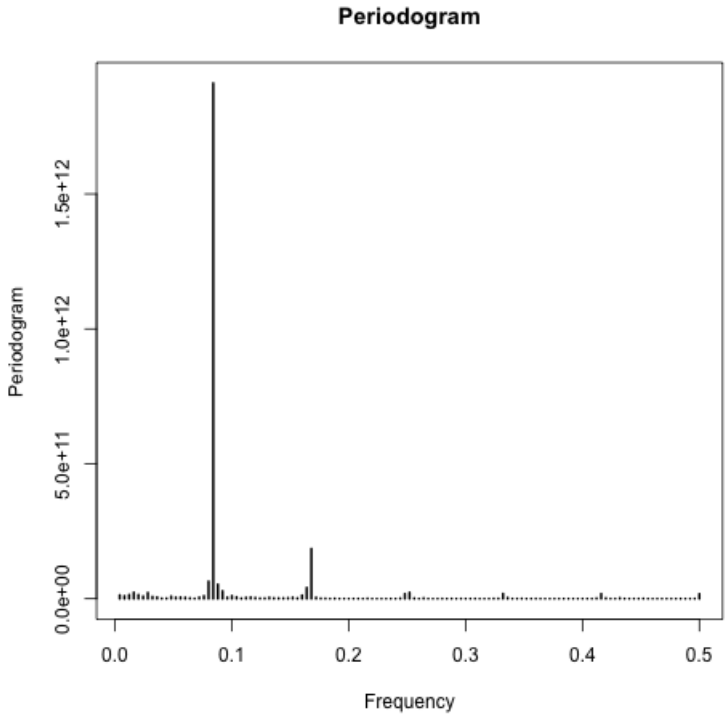
Forecast from the seasonal ARIMA Model

After knowing seasonal ARIMA is a stable estimation of time varying trends and seasonal patterns, I used the forecast model by Frank Davenport with my best fitted seasonal ARIMA model to forecast. I set the 17 years of my data into training set and forecast the rest of the 45 months. The red line indicates the actual data, while the black line is the forecast. Although the forecast is underestimating the actual data points, it did a good job of estimating it within 95 % confidence intervals.

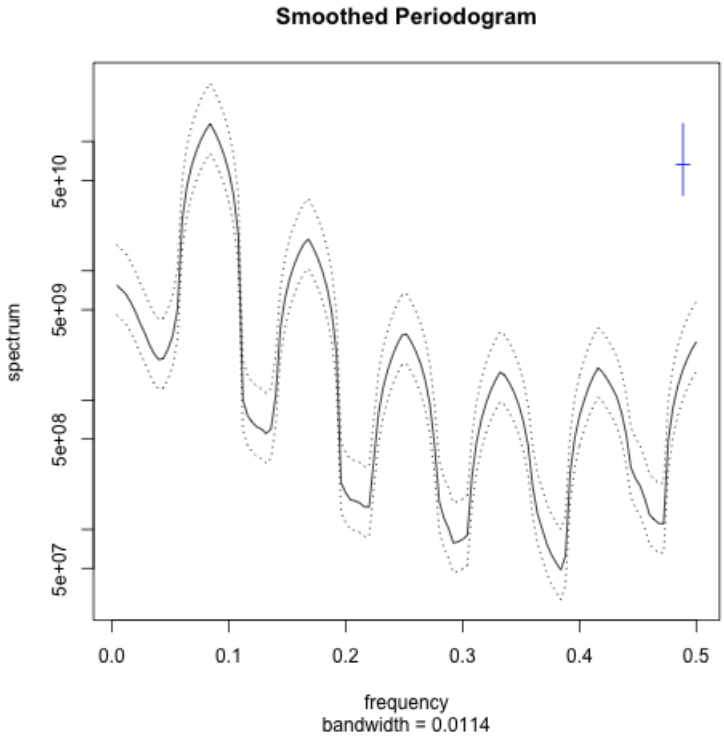


Spectrual Analysis

Now, I moved onto frequency domain analysis. First, I looked at the periodoram to estimate the spectral density. The time series appears to be smooth because there are more values of low frequencies. The highest peak of the periodogram is during the frequency of 0.084.



The basic periodogram is a spectral estimate with high variance. The smoothed periodogram is a better spectrum estimate because the method uses a Fourier transform of the time series itself. The highest spectrum also take place in frequency of 0.08, vaguely followed by 0.16.



Conclusion

After experimenting frequency domain and time domain methods, I found that seasonal ARIMA of (2, 0, 1) X (0, 1, 2)[12] is a good fit for my dataset. I carefully selected the p, d and q values for my ARIMA model based on the lowest AIC and auto aroma estimation. By plotting the time series with the original data, I can easily tell that there are peak periods during June, July and August due to summer vacation. There is also a smaller peak in December for winter vacation. From the seasonal decomposition of Time Series by loess, it shows there is a seasonal pattern and an increasing trend. I detected a few outliers from my datasets. These outliers are caused by 9/11 attack, SARS virus and the depreciation of Canadian dollar. From the coefficients of the GARCH model and the McLeod Li test, I found out that the GARCH model is unnecessary. Next, I used the forecast model by Frank Davenport with my best fitted seasonal ARIMA to perform prediction. The result of the

forecast stays within the intervals, which indicates that it is a good forecast model. The answer of my question is that there exists a seasonal patterns for U.S. citizens to travel to Canada by air based on the seasonal ARIMA model.

References

1. "8 ARIMA Models." 8 ARIMA Models | OTexts. Online Open Access Textbooks. https://www.otexts.org/fpp/8/.
2. Hong, Johnny. STAT 153 Lab 10. http://jcyhong.github.io/stat153_lab10.html
3. "Monthly Tourism Statistics - U.S. Travelers Overseas." Monthly Tourism Statistics - U.S. Travelers Overseas. http://travel.trade.gov/research/monthly/departures/index.html
4. Ozkan, I. ARCH-GARCH Example with R. http://yunus.hacettepe.edu.tr/~iozkan/eco665/archgarch.html
5. "Using R for Time Series Analysis." Time Series 0.2 Documentation. http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html