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

Leanne Lee

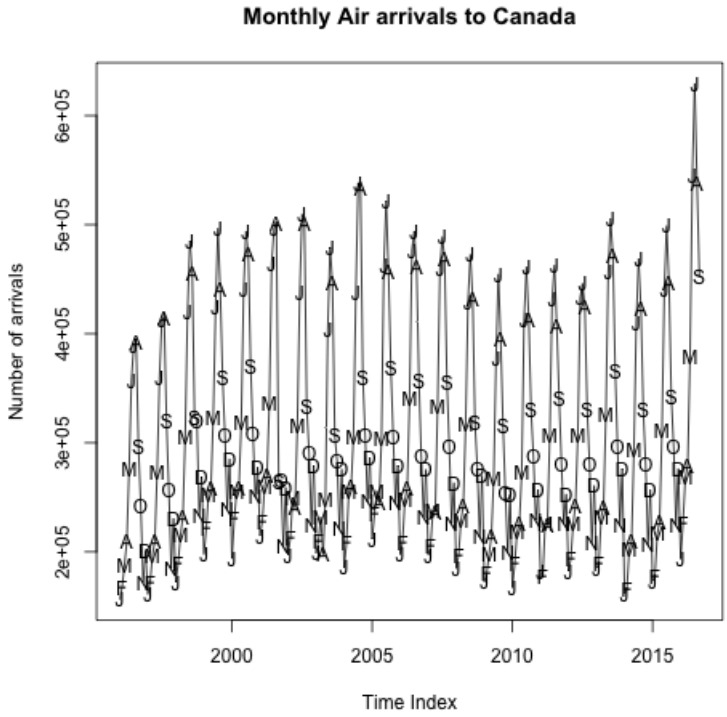
Introducion

The scientific question motivating my work is "is there a seasonal trend for U.S.citizens to travel to Canada by air?" Canada is one of the closest country across the border of the United States. Many U.S. citizen consider Canada a great spot to travel during holiday. In the West Coast, we neighbor near Vancouver and Victoria, while the east coast has Toronto, Montreal and Quebec.

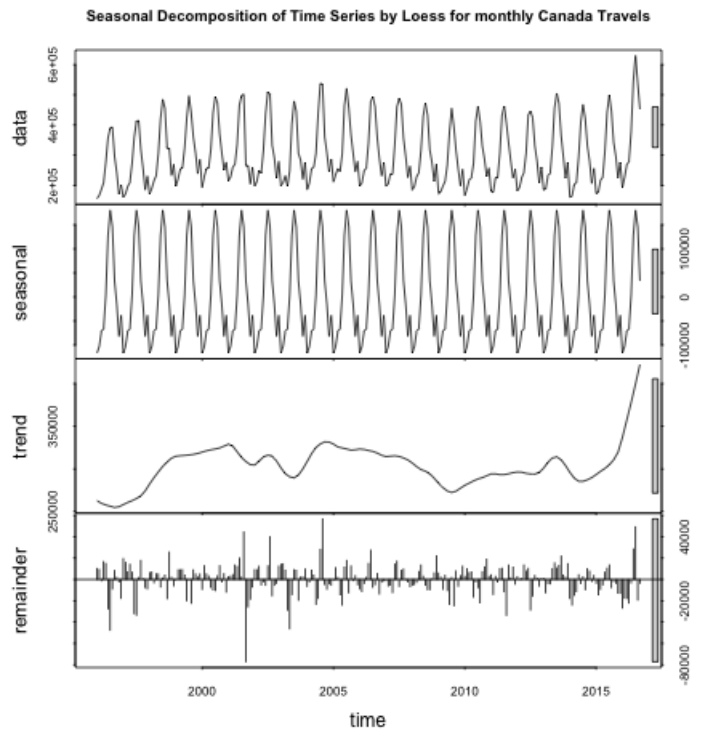
Dataset

The dataset comes from the U.S. Department of Commerce, National Travel & Tourism Office. [NTTO Monthly Tourism Statistics] (http://travel.trade.gov/research/monthly/departures/index.html). The dataset contain U.S. outbound travel by world regions from Jan 1996 to Sep 2016. There are 249 monthly data points throughout these twenty years.

Exploratory Data Analysis



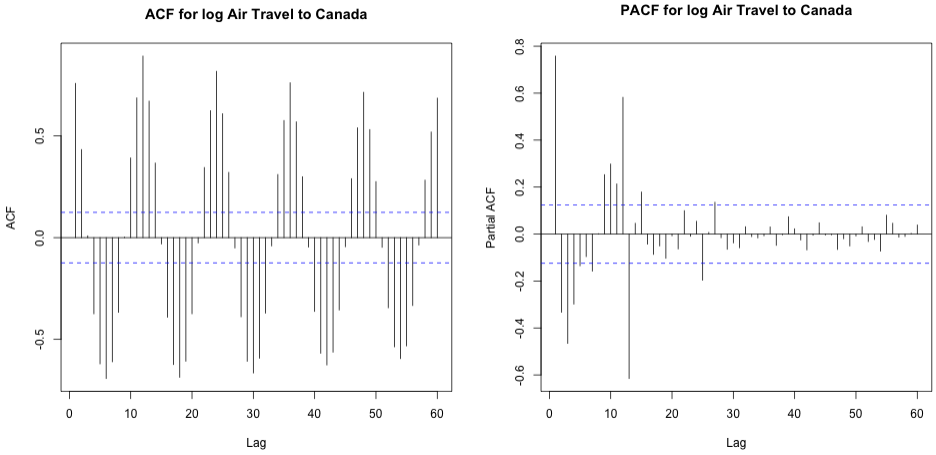
First, I plot the time series for the monthly air arrivals from U.S. to Canada. By examining the time series, the plot looks seasonal with high peak in June, July and August. Students usually travel during their summer break and this is the main drive of the increase in air travel in these months. There is also a smaller peak in between of each cycle. The small peaks take place in December, which is during Christmas break and students tend to go travel more often. Because the graph doesn't get effect when the variation increase with the level of the series, so I didn't consider the log of the time series.



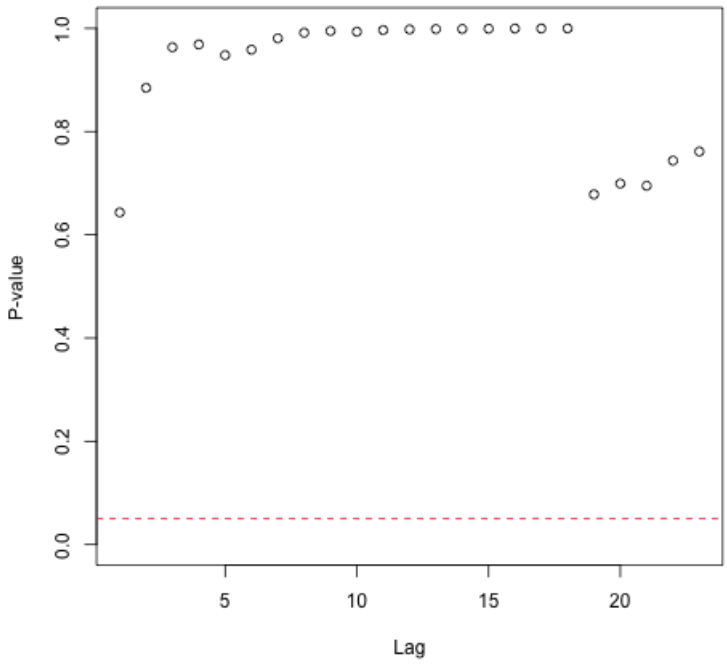
After the basic inspection on the time series, I ran the **stl()** command for a season decomposition of Time Series by Loess. I first inspect the data section. The data looks stationary. I consider the bar on the right hand side as one unit of variation. The bar on the seasonal panel is slightly bigger 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 the trend is not dominating, but there's an upward trend at the end. The spike upward trend of air traveling to Canada in this summer is because the low Canadian dollor, causing a boost to tourism in Canada. The remainder panel shows there are some high residuals in year of 2002 - year of 2005 and year of 2016.

ARIMA Model

ACF and PACF

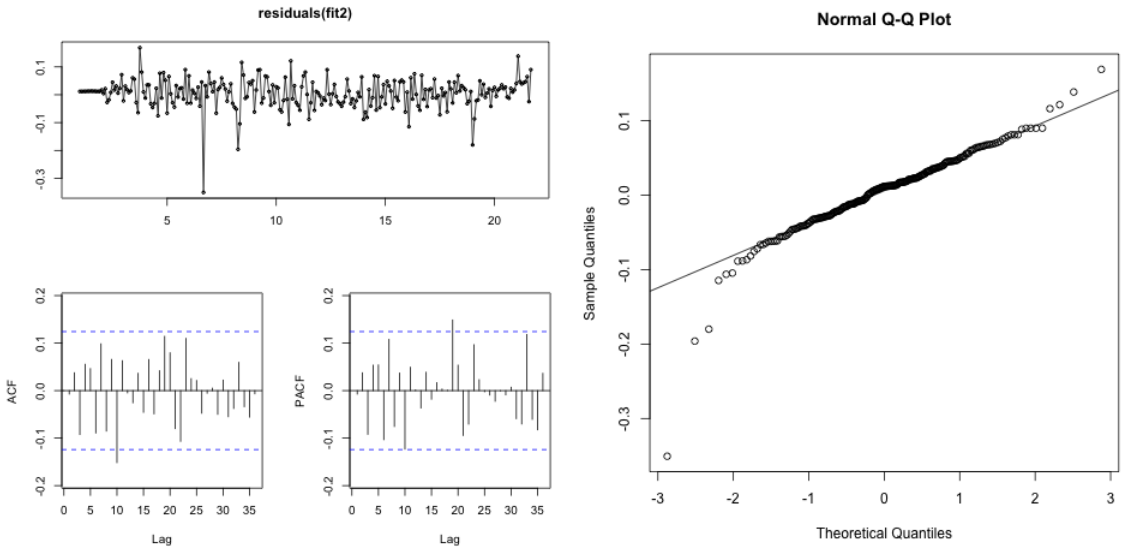


Based on the results on ACF and PACF, I found that there are seasonal autoregressive and moving average effects. So I ran the ARIMA model with changing order of p,d, q and seasonal orders. First, I ran the auto arima to get a rough estimate of my orders of seasonal ARIMA. Since the original series looks stationary, I do not need to consider the high order of differencing(**d**). 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 $(1,1,1)X(0,1,2)_{12}$ Since the seasonal pattern is strong and stable, I use (0,1,2) as a seasonal difference. The best fit ARIMA model with the lowest AIC is **ARIMA** $(2,0,1)X(0,1,2)_{12}$



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.

Residuals of the Model



After fitting the model, it is important to diagnostic the residuals of the ARIMA model. The

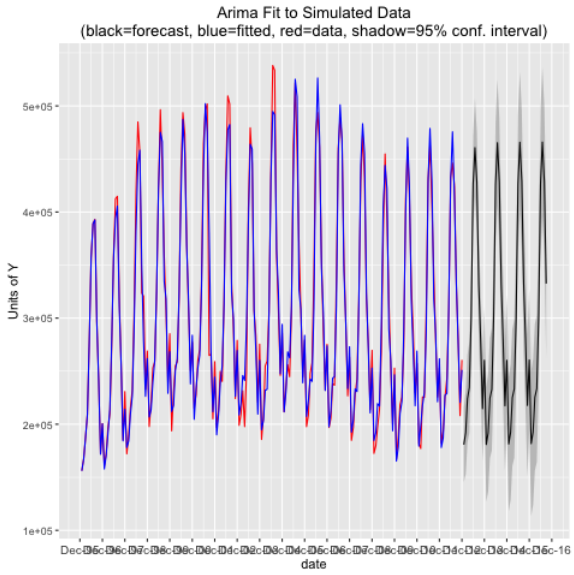
blue line indicates the 95% confidence interval under the null hypothesis of white noise. Both ACF and PACF show the majority of residuals are under the blue thereshold and it passes the residual test. The residuals of fitted models has a few outliers in 2002, 2004 and 2014. I also exaimined the QQ plot to check the normality of the residuals. The plot shows there are a few outliers in the beginning. The most significant one is the first point, which causes by the extreme high value during July 2016.

Detecting Outliers

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

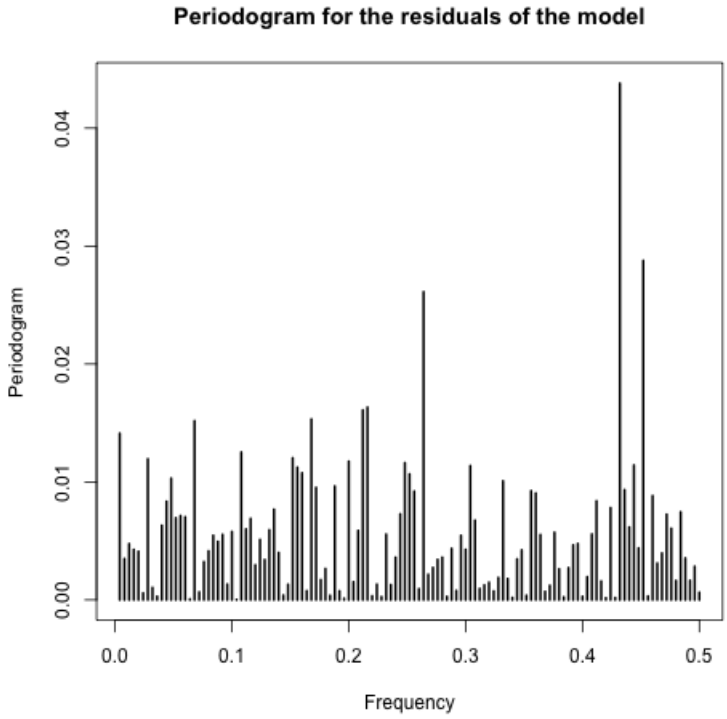
From the residuals plot, I decided to use **detectAO()** and **detectIO()** to determine the outliers. There is no result for detecting additive outliers (AO), so I exaimed the innovative outliers (IO) and found 3 outlier in index 69, 88 and 217. The biggest magnitude comes from index 69, which is September 2001. The main reason of the decreased in travelling to Canada is because of the 9/11 attack. People are being more skeptical of taking air flights in September 2001.

Forecast from the ARIMA Model

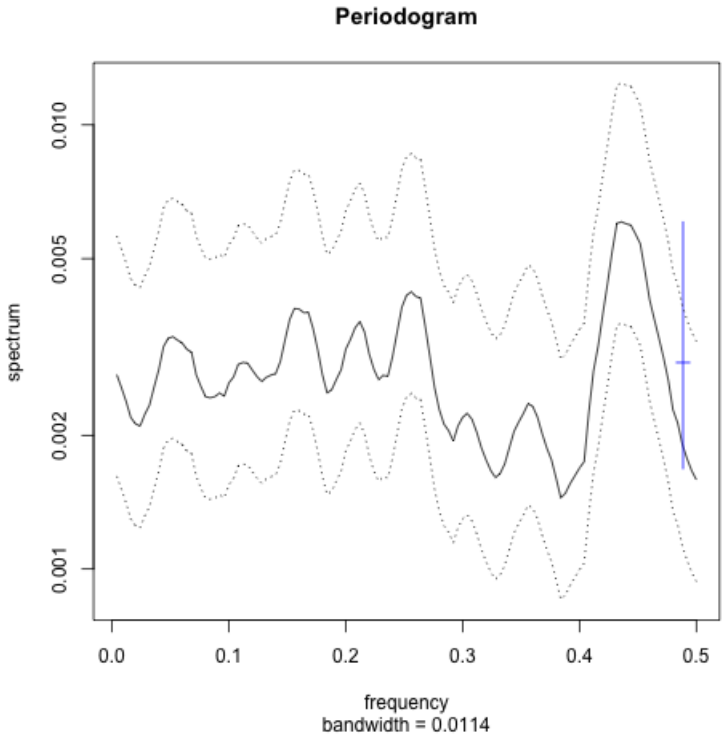


After knowing seasonal ARIMA is a stable estimation of time varing trends and seasonal patterns, I used the forecast model by Frank Davenport to forecast using my best fitted ARIMA model. I set the 17 years of my data into training set and forecast the rest of the 45 months. The forecast is underestimating the actual data points.

Spectrual Analysis



The time series appears to be smooth because there are more values of low frequencies. There are only a few peaks in this periodogram in frequency 0.27, 0.43 and 0.46, where the time series has a strong sinusoidal signal for these frequency.



I can draw a flat line from the smoothed periodogram between the intervals; thus, there is no ambigity. The periodogram demonstrates very stable.

ARCH-GARCH Model

Although I used McLeod-Li test to check if there is a need of Arch-garch test, I want to reassure my assumption by running the garch model. 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).

```
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
```

Conclusion

The answer to my question is ..

References

- http://jcyhong.github.io/stat153_lab10.html
- https://www.otexts.org/fpp/8/
- http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
- http://yunus.hacettepe.edu.tr/~iozkan/eco665/archgarch.html
- http://travel.trade.gov/research/monthly/departures/index.html