Time Series Analysis of the Air Pollution in Beijing

M.S. Statistics Candidate - 2019

Bruce Zhu (UCLA)

2018.03

Background

Motivation :

Beijing is my hometown. It used to be a very beautiful city when I was a child; however, due to the developments of industry, the air pollution in Beijing become more and more serious , which affect the health of the residences including my family as well as many of my friends. Therefore, I hope I can discover some patterns from the historical data of PM2.5 concentration (ug/m^3) in Beijing so that I am able to provide some useful advices on air pollution for those people I love.

What is PM 2.5 :

Fine particulate matter (PM2.5) is an air pollutant that is a concern for people's health when levels in air are high. PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated.

Used to be…

Now…

Data

Data Source :

I gather the data from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data), which records the PM2.5 concentration in Beijing every one hour from Jan 2011 to Dec 2014. The original dataset has 13 variables (including PM2.5 concentration) and 43824 records.

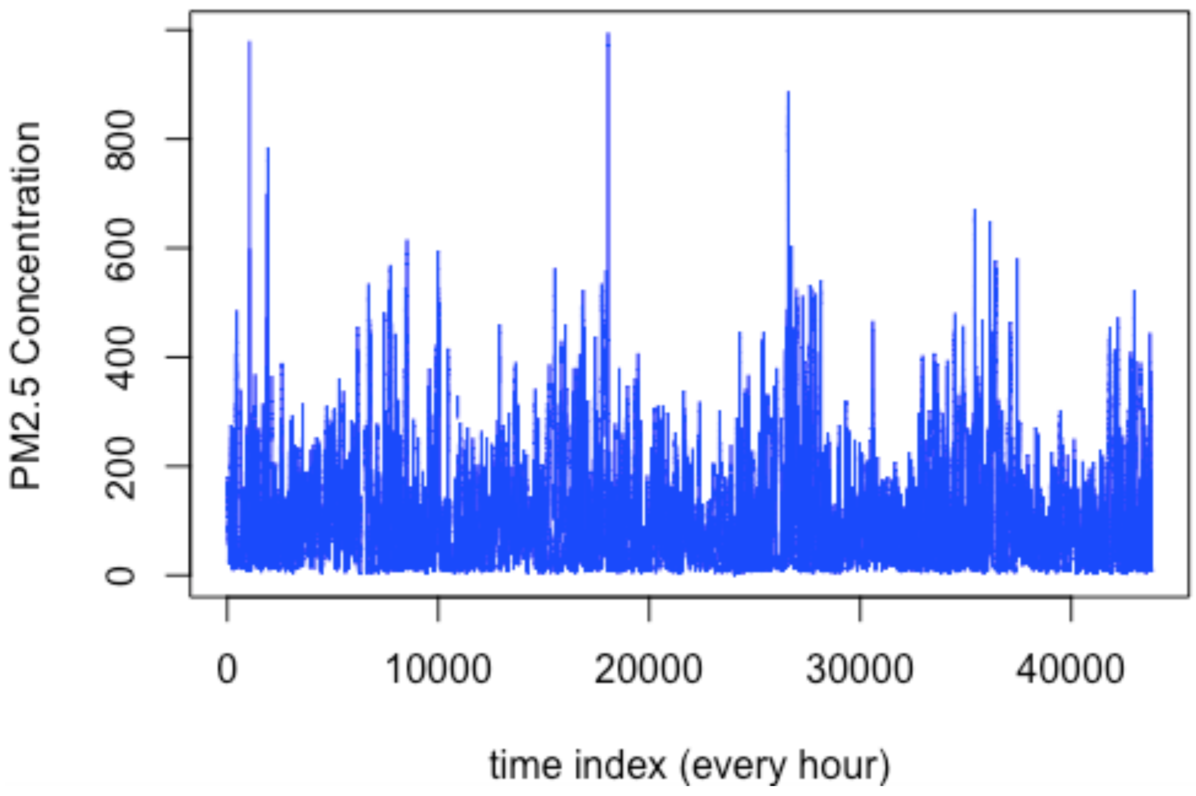
Data Preprocessing :

Since 43824 records are not convenient for visualization. I calculated the mean of each day from Jan 2013 to Dec 2014 as my new records (i.e. daily records of 2013 as well as 2014), then I got 730 records of PM2.5 concentration as my analyzing dataset.

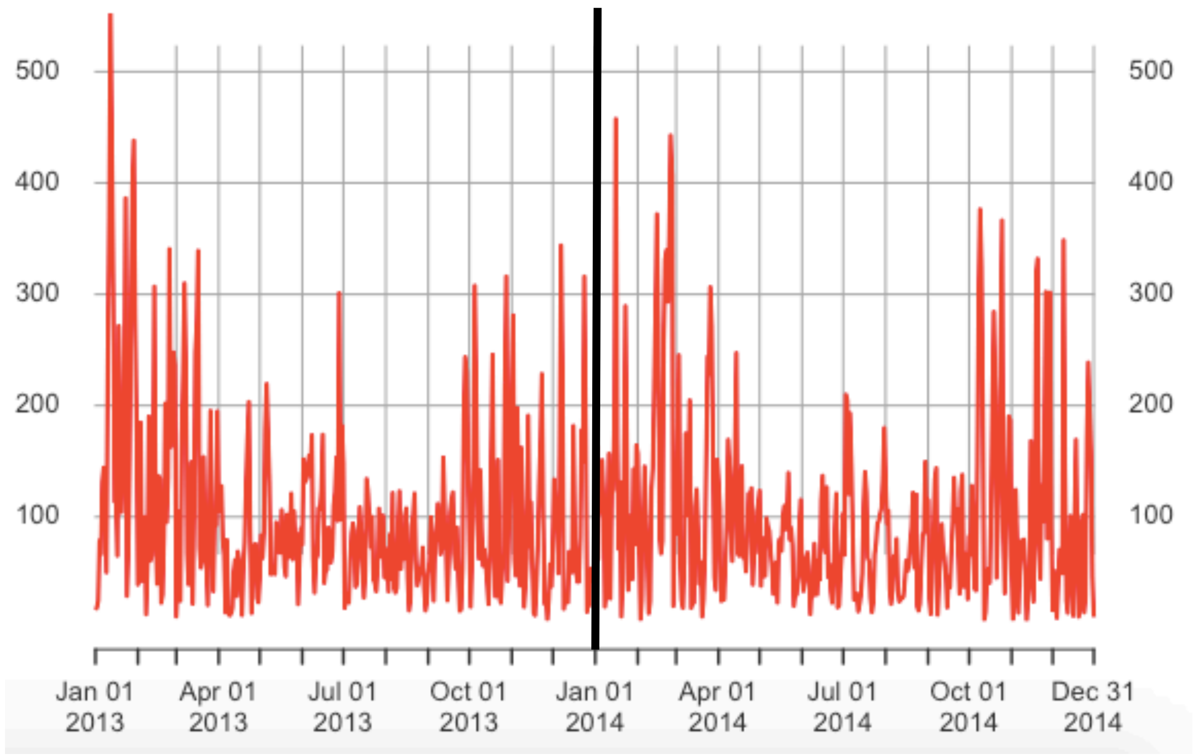
Original Data V.S. Preprocessed Data

- Using processed data is convenient for visualization; thus, I choose to use preprocessed dataset to conduct the following time series analysis. Besides, we can see also see there should be some small cycles in both datasets.

PM2.5 Concentration of Original Data

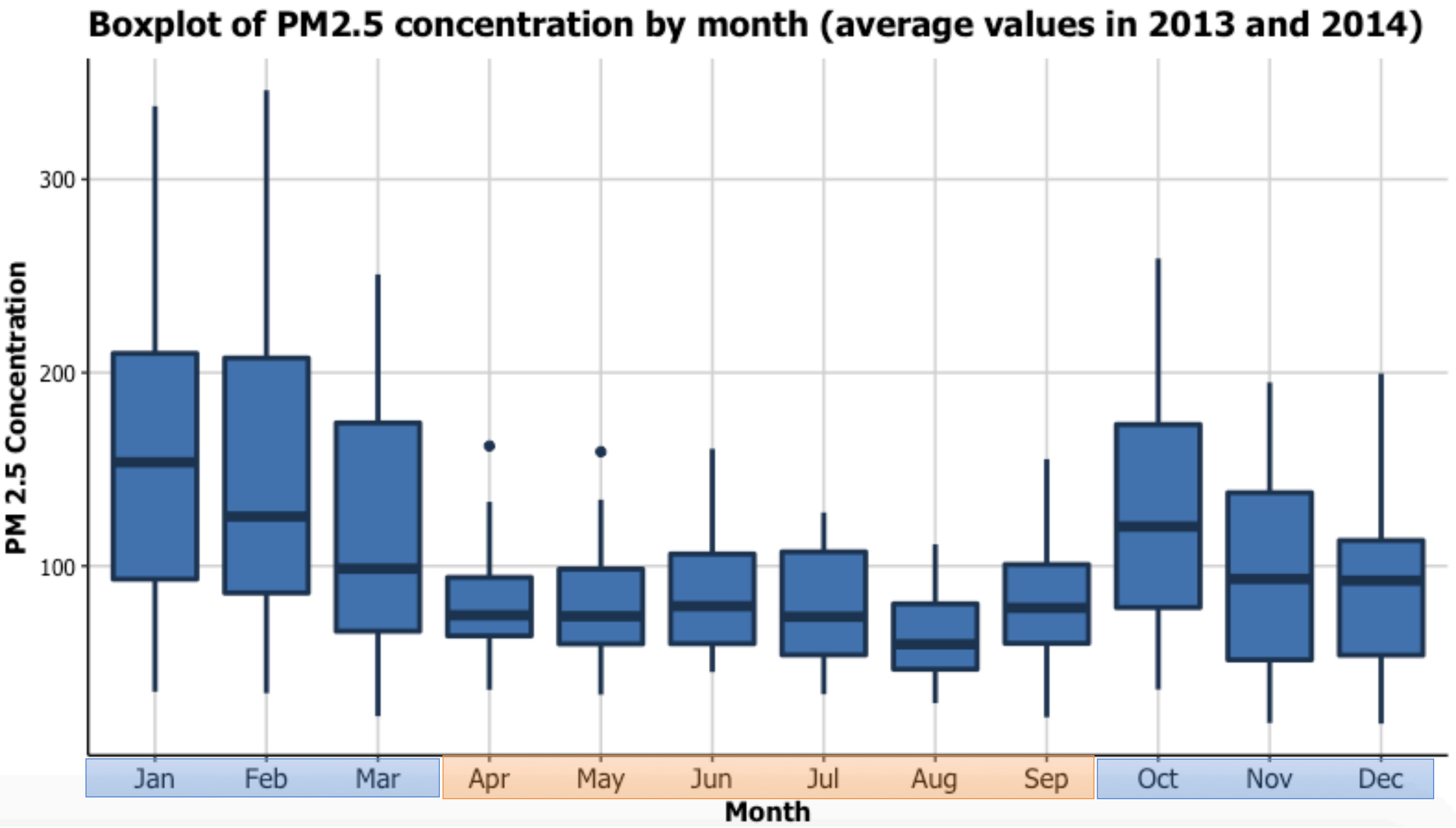


PM2.5 Concentration of Processed Data 2013-01-01 / 2014-12-31



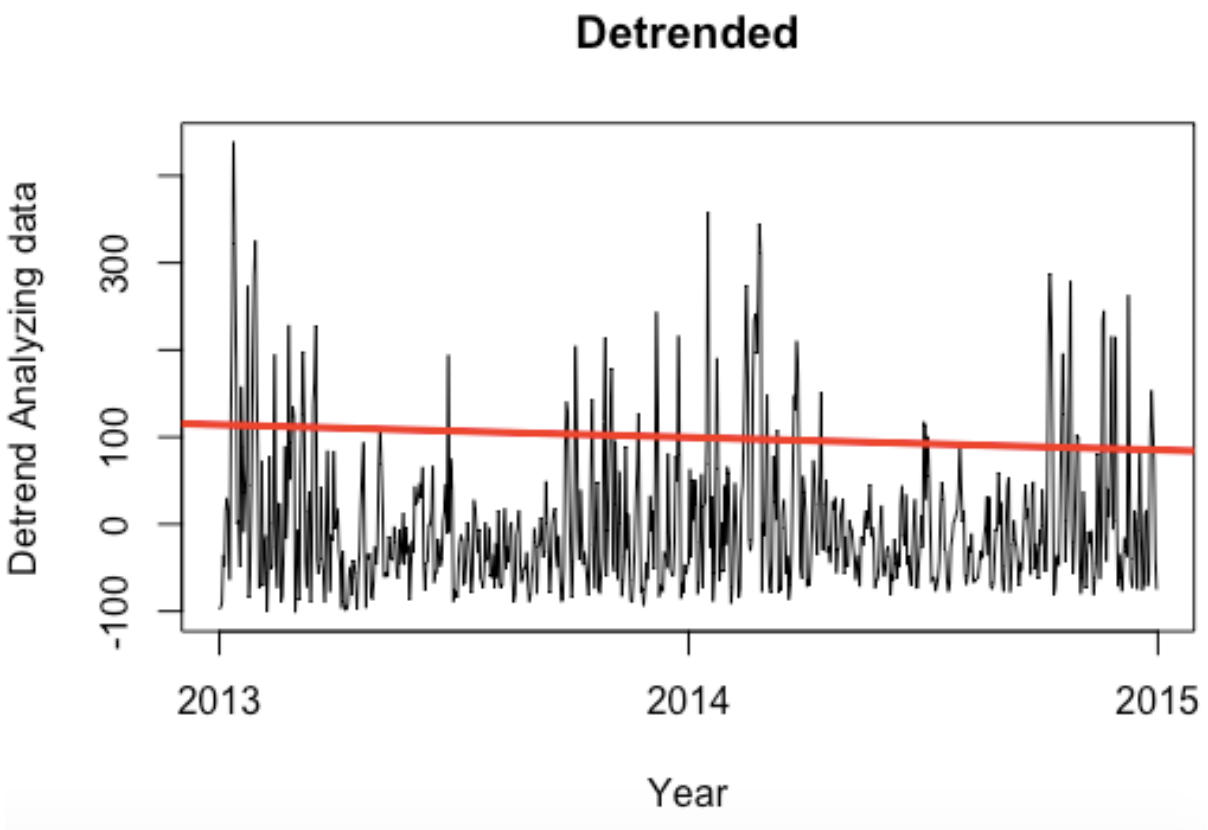
Box Plot of PM 2.5 Concentration by Month

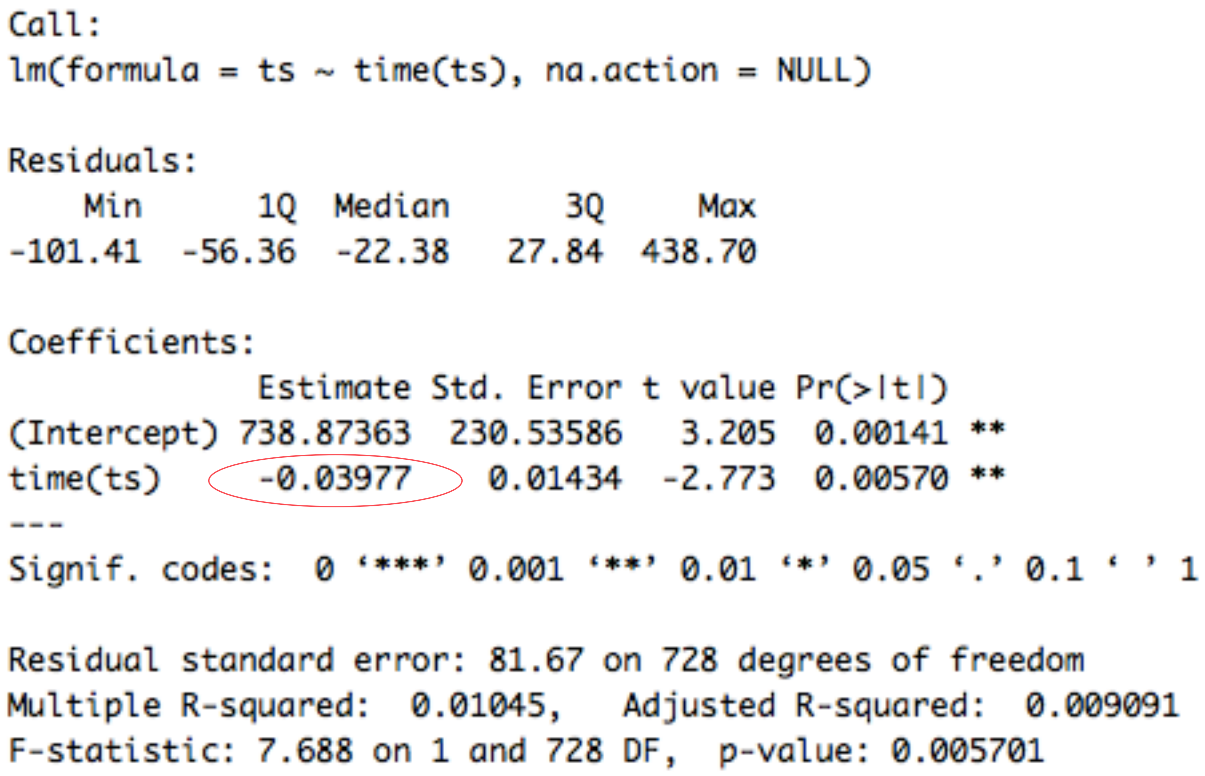
- From October to March (cold), the mean and variance of PM 2.5 is relative large, and from April to August (hot), the concentration of PM2.5 is relative small and stable, which we may be caused by the temperature.



Trend Detection

- There is a very weak decreasing trend in the dataset; however, the slope is close to 0; hence, should not strong enough to dominate. Considering the model complexity, I choose not to differentiate the data.

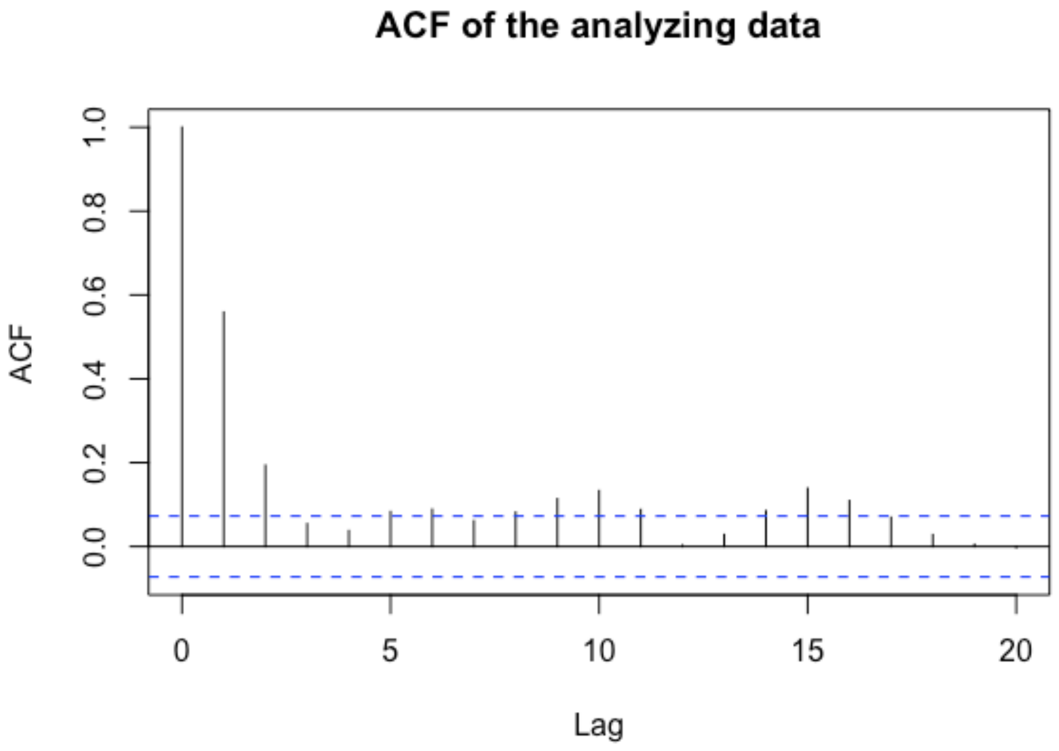




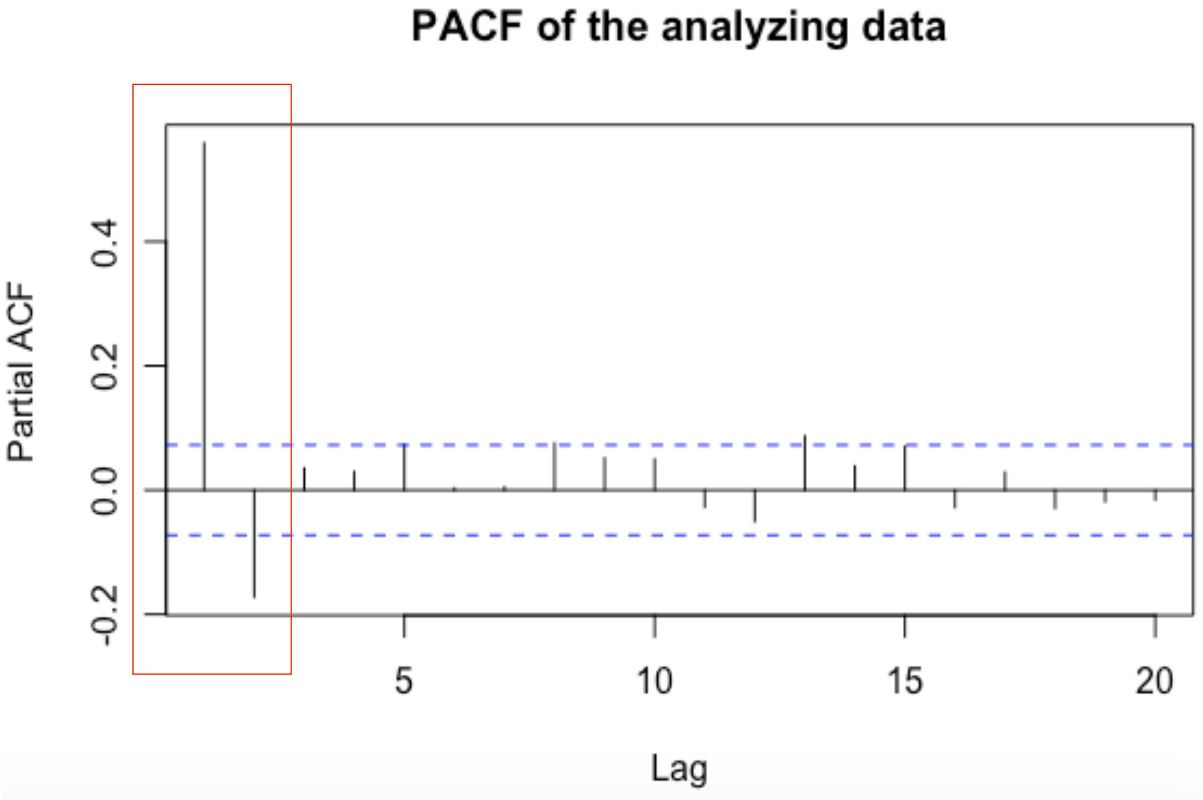
```
Call:
lm(formula = ts ~ time(ts), na.action = NULL)

Residuals:
    Min      1Q  Median      3Q     Max
-101.41  -56.36  -22.38   27.84  438.70

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 738.87363  230.53586   3.205  0.00141 **
time(ts)     -0.03977    0.01434  -2.773  0.00570 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.67 on 728 degrees of freedom
Multiple R-squared:  0.01045,   Adjusted R-squared:  0.009091
F-statistic: 7.688 on 1 and 728 DF,  p-value: 0.005701
```

ACF/PACF

- From the ACF graph, we know that our analyzing data should not have very strong trend; thus, differencing may be not necessary.
- Since the PACF cuts off after Lag 2, the data might fit an AR(2); however, we should try fitting different models so that we can get the best.



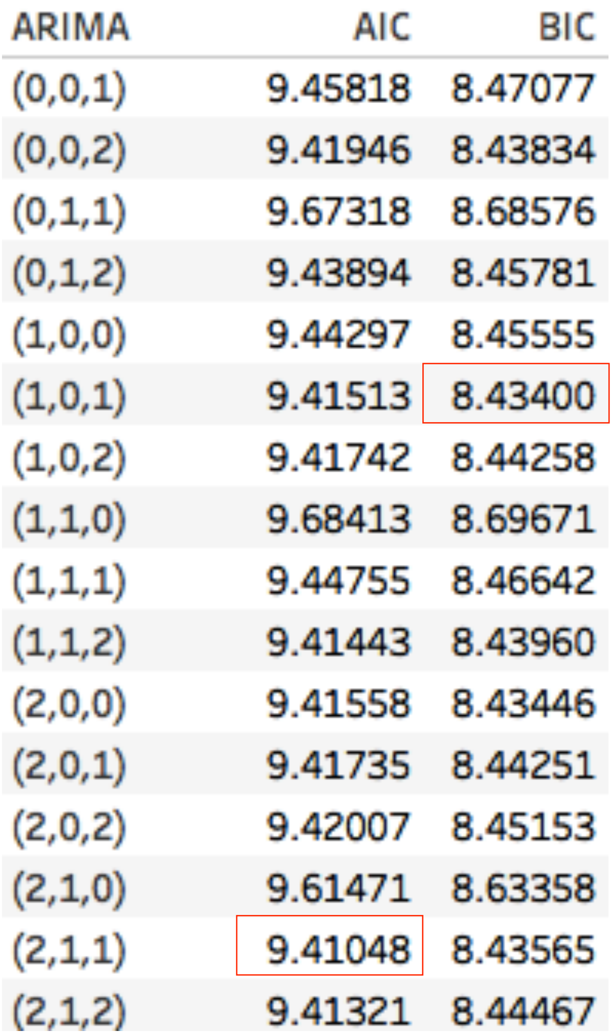
Seems Tails off



Seems cutting off after lag 2

Model Fitting

- Grid search different parameters around 2 to fit ARIMA model. Under AIC criterion we should choose ARIMA(2,1,1), and we should choose ARIMA(1,0,1) under BIC. Considering the complexity of model, I choose model ARIMA(1,0,1) under BIC criterion.

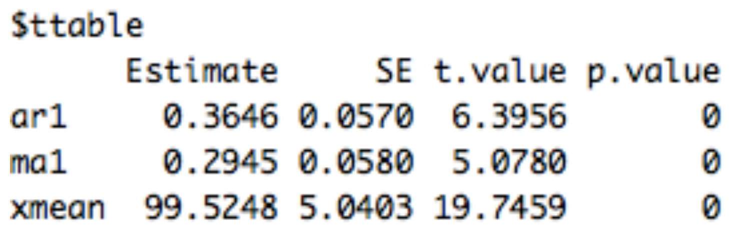


ARIMA	AIC	BIC
(0,0,1)	9.45818	8.47077
(0,0,2)	9.41946	8.43834
(0,1,1)	9.67318	8.68576
(0,1,2)	9.43894	8.45781
(1,0,0)	9.44297	8.45555
(1,0,1)	9.41513	8.43400
(1,0,2)	9.41742	8.44258
(1,1,0)	9.68413	8.69671
(1,1,1)	9.44755	8.46642
(1,1,2)	9.41443	8.43960
(2,0,0)	9.41558	8.43446
(2,0,1)	9.41735	8.44251
(2,0,2)	9.42007	8.45153
(2,1,0)	9.61471	8.63358
(2,1,1)	9.41048	8.43565
(2,1,2)	9.41321	8.44467

ARIMA(1,0,1)

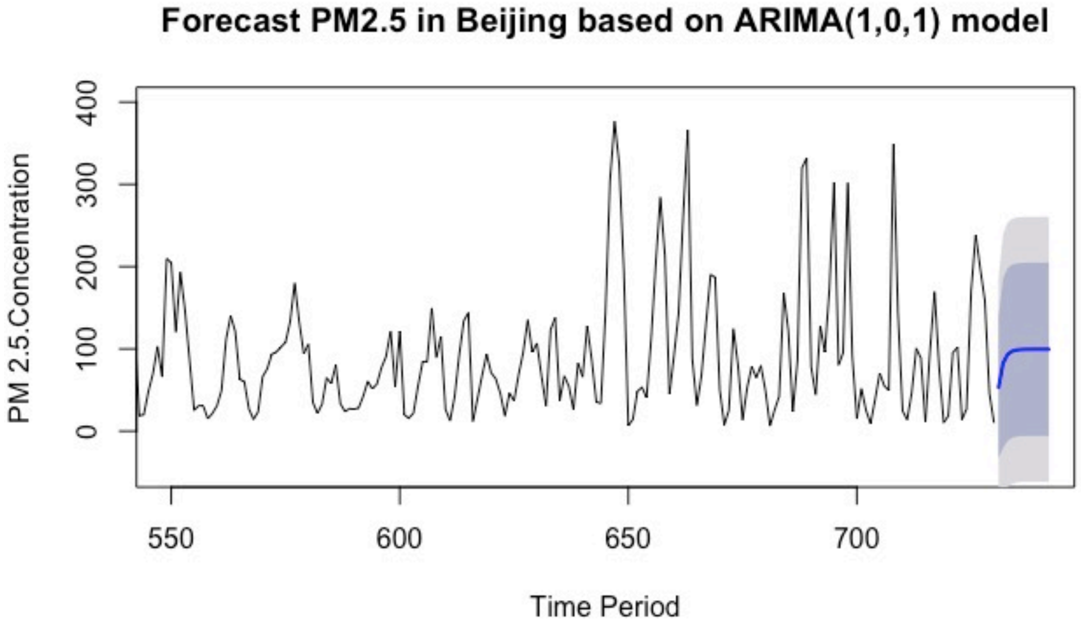
BIC Criterion

Estimated Model ARIMA(1,0,1)



```
$ttable
      Estimate     SE t.value p.value
ar1     0.3646 0.0570  6.3956       0
ma1     0.2945 0.0580  5.0780       0
xmean  99.5248 5.0403 19.7459       0
```

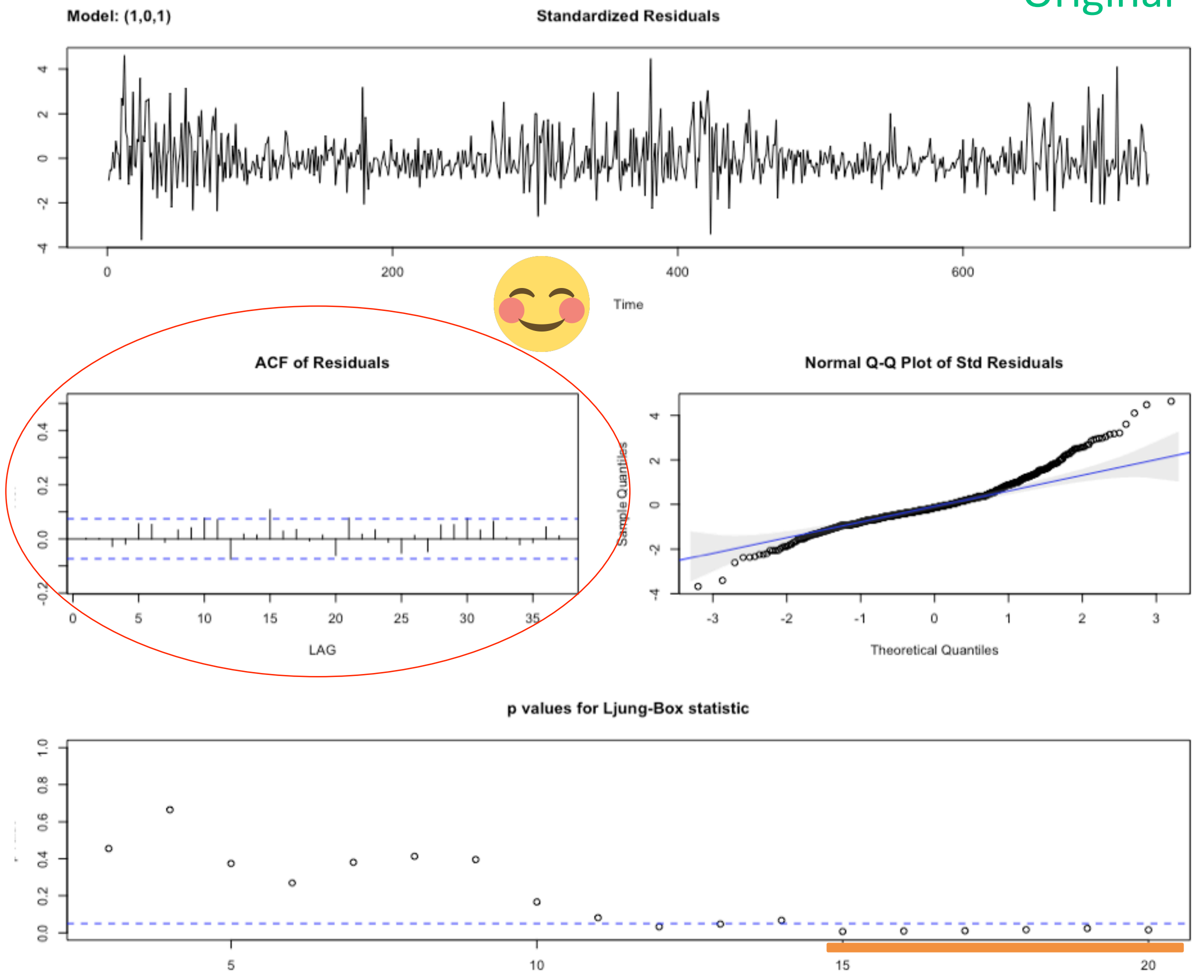
Prediction Based on Model ARIMA(1,0,1)



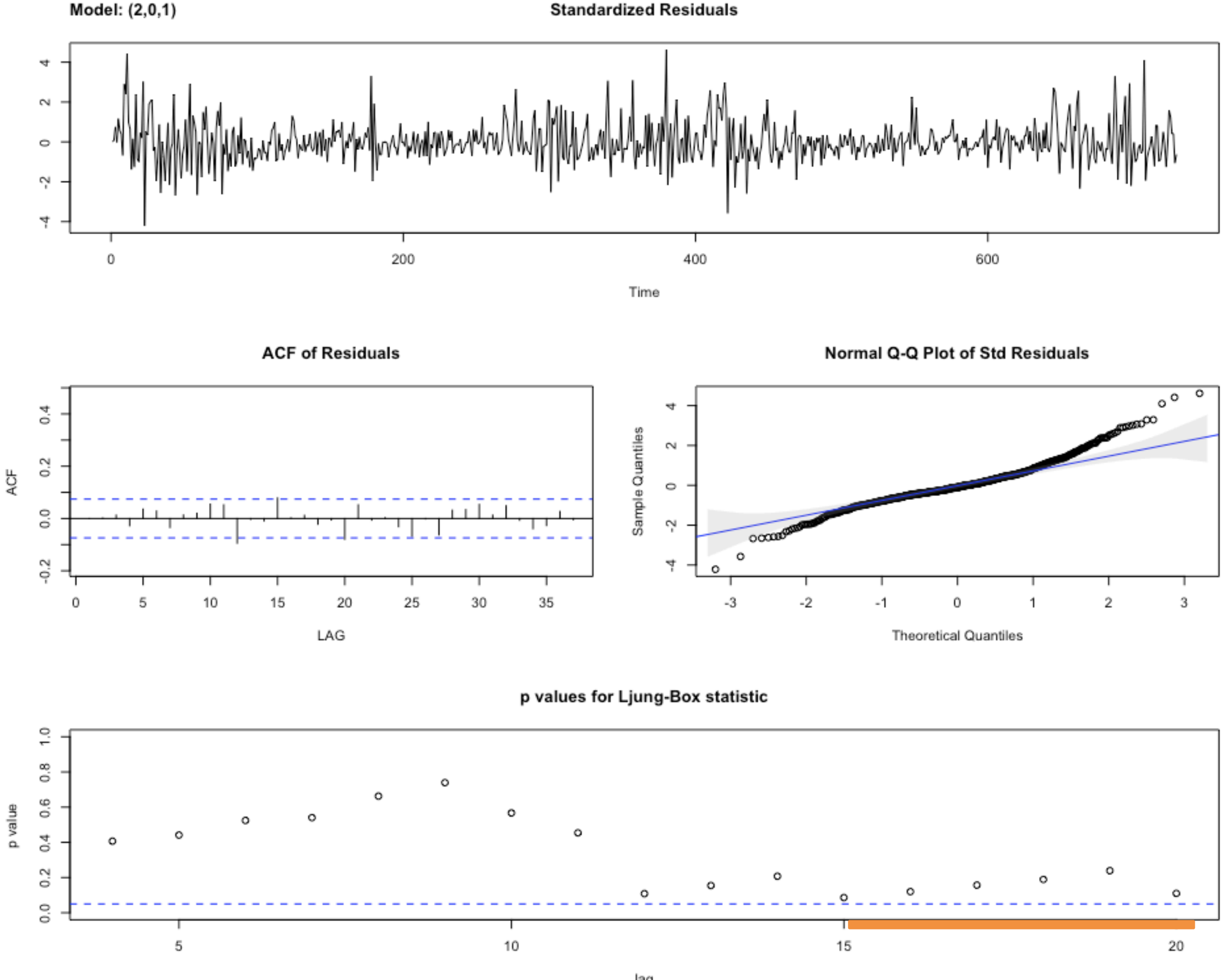
Model Diagnostics

- The left graph is the diagnostics of ARIMA(1,0,1) I built above, which is good. But I found the p-value after 15 are below 0.05; thus, I tried to conduct differentiation for the data, and fit the best model ARIMA(2,0,1) for differentiated data. Although, the p-value after 15 very smaller above 0.05, I still choose to use the ARIMA (1,0,1) for lower model complexity.

Original

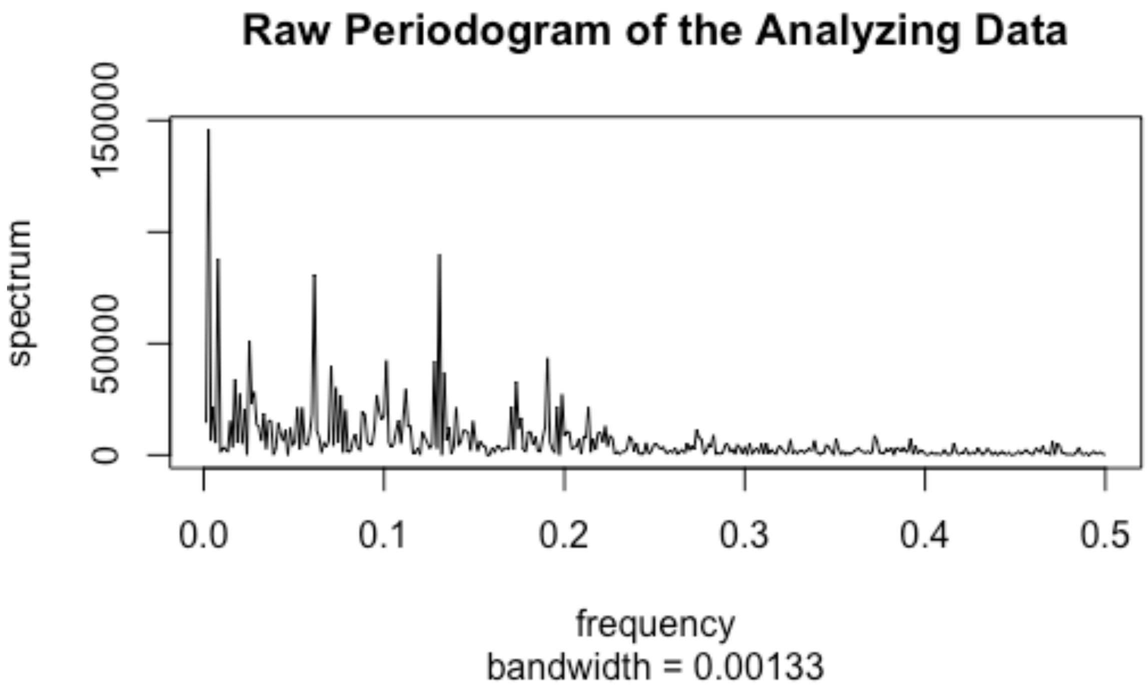


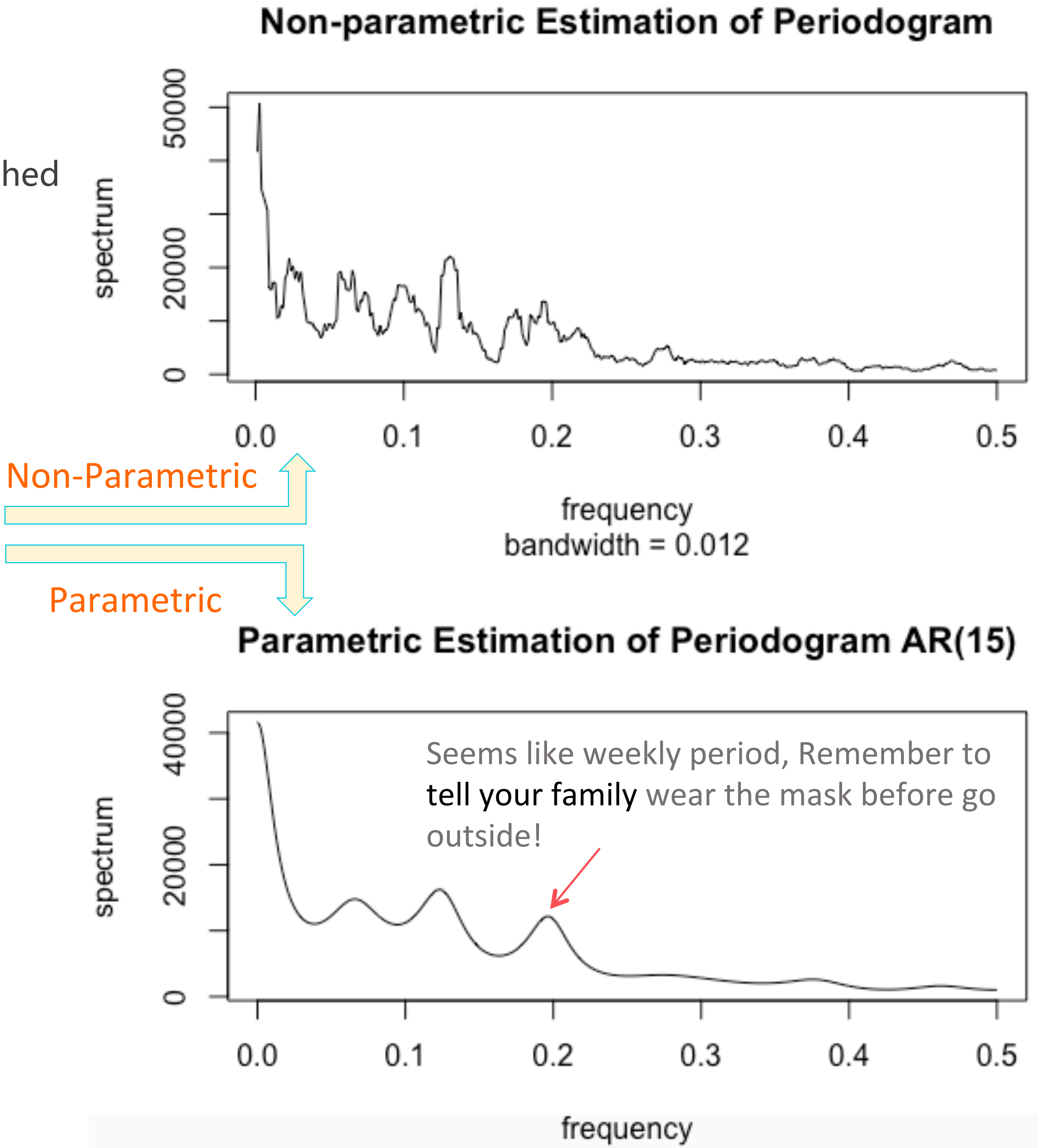
Differentiated



Frequency Analysis

- Both Non-parametric and parametric estimation smoothed the raw periodgram well, and we can see there are four main peaks in the graph.

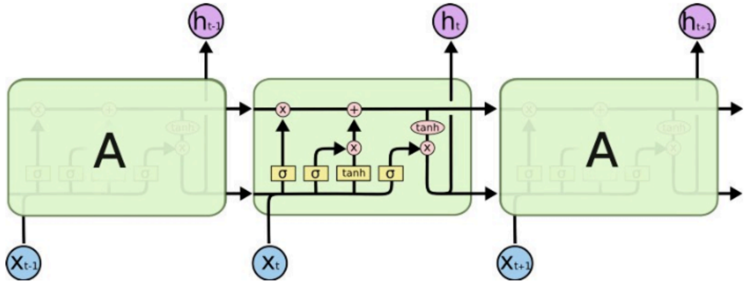




Seems like weekly period, Remember to tell your family wear the mask before go outside!

Next

1. Try using Deep Learning (LSTM) model to predict :



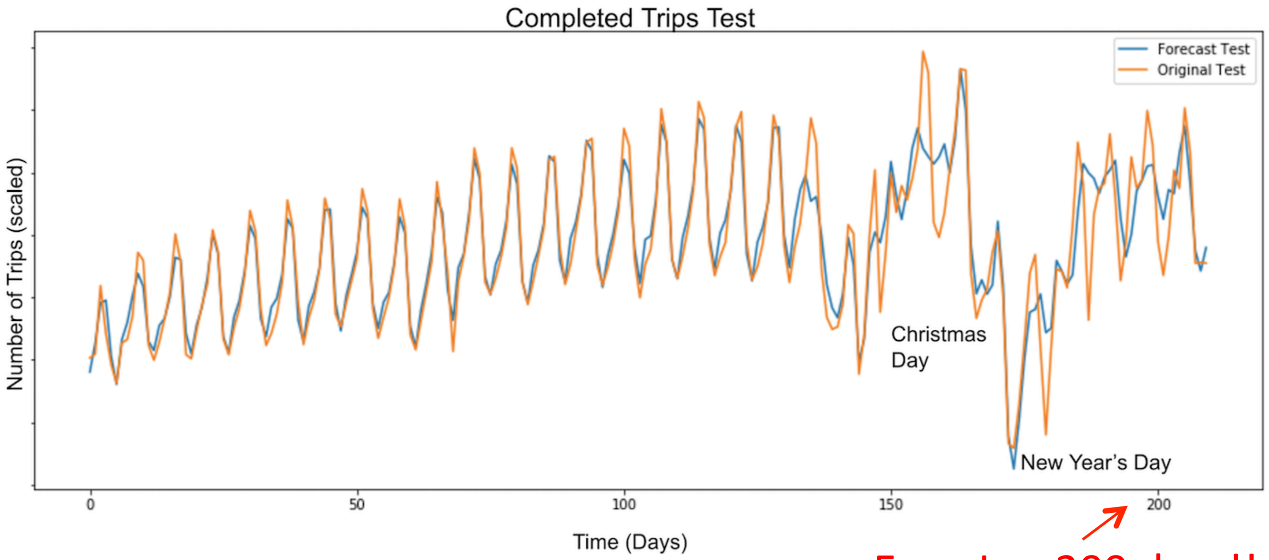
LSTM Architecture

Pros:

- Can find out very complex relationship among data and improve the accuracy of prediction.
- Not necessary to determine many parameters subjectively.

Cons:

- Hard to explain.
- Hard to train (Need more data to converge)



Even Lag 200 days!!

2. Find more recent data and more variables:

The dataset I found only end up in 2014, we need more recent data to find out the current trend;
Besides, we can also try to see if other variables have effects of the PM2.5.

Note: Uber AI lab have successfully used LSTM on Time Series forecasting： https://eng.uber.com/neural-networks/

Thanks