Analysis of Air Pollution in Beijing

Bruce Zhu
University of California, Los Angeles (UCLA)

Mar 13, 2018

Abstract

Due to the rapid developments of industry, air pollution in Beijing become more and more severe in recent years. The harmness caused by this pollution has a significant impact on the health and daily life of the Beijing residents. Hence, it should be meaningful if we could discover some patterns from the air pollution data, so that the conclusions can be used to protect the Beijing residents in advance. In this paper, the PM2.5 data in Beijing from year 2013 to 2014 is analyzed. We fit an ARIMA(1,0,1) model to identify the relationship among the data as well as do a 5-day ahead prediction; besides, Spectral Analysis is also conducted to estimate the possible cycles for the air pollution in Beijing.

1. Introduction

Beijing is used to be a very beautiful city. However, due to the rapid development of industry. it become a representation of air pollution right now. For the residents living in Beijing, masks become their life necessities, which they have to wear to protect their health.

From the report on U.S News, "Air pollution can cause lung disease and cardiovascular disease, as well as cancer and birth defects", indicating the harmness of air pollution. Besides, the report also points out "such high levels of air pollution are a relatively recent phenomenon, and long-term health consequences take time to develop. Most studies about air pollution-triggered health effects have been conducted in environments with much less pollution than exists in China today"[1], which means the long-term effect of peoples' health in Bei-

jing might turn out to be worse than we imagine. Therefore, it's necessary to conduct analysis on PM2.5 in Beijing so that people can prevent in advance.

The goal of this project is to analyze the trend and periodicity of the PM2.5 concentration in Beijing. In addition, fitting an ARIMA model to do some predictions.

2. Dataset

The dataset is downloaded from UCI Machine Learning Repository[2]. The raw data contains 43824 hourly-recorded instances from 01/01/2010 to 12/31/2014. For the convenience of visualization as well as filling the missing values, we calculate the mean of daily concentration of PM2.5 from 01/01/2013 to 12/31/2014; hence, we change our data from hourly-recorded to daily-recorded, which leave us 730 observations. We are going to use this transformed dataset as our analyzing data to conduct the following analysis.

3. Data Exploration

Compared with Figure1 and Figure2, we can see that our analyzing dataset is more clear to visualize. The mean of our analyzing data is 99.72 and its median is 77.40. However, from the PM2.5 AQI conversion scale table shown in Figure16[3], both the mean and median of analyzing data belongs to the category of Unhealthy, which proves again the air pollution are very severe in Beijing.

Besides, It is easy to notice that the concentration of PM2.5 varies in different months from Figure2. In order to detect this variance, we draw boxplot of PM2.5 concentration as shown in Figure3. It is interesting that the concentration of PM2.5 seems relative low and stable in high-temperature months and reversed in low-temperature months, which might be caused by the strong effect of convection under high-temperature. Therefore, summer should be a recommended season for passengers to visit Beijing, which is good for their health.

After having a big picture of the analyzing data, we explore the data from the perspective of time series. The PM2.5 data is regressed on time t as shown in Figure4, and the result of summary in Figure5 tells us that there exists a very weak trend in our dataset. Since the slope of regression line is only -0.03977 (which is very close to 0), that slope seems very hard to become dominant in our dataset. Under the real industrial environment, balancing the model complexity and performance is always an very important thing to consider. Hence, for model complexity, not differentiating the data will be a good choice.

Besides, Figure6 and Figure7 show the ACF and PACF respectively. The ACF seems tails off; while the PACF shows cutting off after Lag2, which suggest an AR(2) model; however, our subjective judgement are not 100% reliable. Therefore, we need to fit various ARIMA models around AR(2) in the below section to find our best model.

4. Modeling Fitting and Diagnostics

Figure8 shows the candidate ARIMA models and their corresponding AIC and BIC values. According to AIC, ARIMA(2,1,1) has the smallest AIC value (9.41048); however, by BIC, ARIMA(1,0,1) should be chosen because of the smallest BIC value (8.43400). Compared ARIMA(2,1,1) with ARIMA(1,0,1), choosing ARIMA(1,0,1) based on BIC criterion will be a better option, because the complexity of ARIMA(1,0,1) is lower than ARIMA(2,1,1); in addition, ARIMA(1,0,1) indicates again that differentiating seems not necessary for our analyzing data. Hence, ARIMA(1,0,1) is chosen to be the preferred model.

Using the analyzing data to fit ARIMA(1,0,1), we get the estimation of our parameters as shown in Figure9. The p-values are all smaller than 0.05, which means all the estimators are significant; therefore, there is no need to drop any parameters to rebuild models. The fitted model suggests us that the concentration of PM2.5 tomorrow may be positive related to the concentration today.

In order to check the reliability our fitted ARIMA(1,0,1) model, diagnostics

are performed. In Figure10, the plot of standardized residuals mostly indicates that there is no trend among the residuals, and the ACF of residuals shows that there is no significant autocorrelations, which are both good results. However, the bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. At begining, those points are all above the dash blue line, which is a good signal; however, the points start to fall below the dash blue line after lag 15, which means there exists significance among accumulated residual autocorrelation after lag 15. We wonder if differentiate the data will help to solve this problem. After differentiating our analyzing data, we repeat above procedures again and find out the best model fitted the differentiated data is ARIMA(2,0,1). The model diagnostics of this ARIMA(2,0,1) are shown in Figure11. From the p-values for Ljung-Box statistics, It looks like all the points are above the blue dash line; however, after lag 15 they are still pretty close to the bound (a little bit above the blue dash line), which means even differentiating the data, we are also not so confident that after lag 15 the accumulated residual autocorrelations are surely unsignificant. Therefore, insisting the original ARIMA(1,0,1) will be a reasonable choice because of its simplicity and robustness.

Therefore, we choose ARIMA(1,0,1) as our final model, and also conduct prediction as shown in Figure12. Our model predict a upper trend for 5 days, which is reasonable, because our prediction starts from a bottom of an obvious cycle. Besides,Figure12 also shows the 95% confidence interval of our prediction on the graph.

5. Spectral Analysis

From the raw periodogram in Figure13, it is very easy to identify the highest peak at frequency 0.0027, indicating the period is $1/0.0027 = 370.4\ days \approx 1\ year$. Except the highest one (yearly cycle), it seems that there also exist some other periodic peaks in the raw periodogram; however, since the raw period is relatively choppy and have many small spikes, it will be easier to identify the peaks if we could smooth our periodogram.

One approach is the non-parametric estimation of periodogram. Modified Daniell kernel is used during the estimation, and the above part of Figure14 shows the result of non-parametric estimation. Although the non-parametric estimation smooth our raw periodogram in some ways, the peaks are serrated, which is still hard for us to determine the cycle.

Since the above method does not give us an ideal result, we may apply the parameteric estimation approach to smooth the raw periodogram, which fit an AR model to the data and use the spectral density of that AR model as an approximation. The bottom part of Figure14 shows the parameteric estimation result, and there are clearly another three peaks except the yearly cycle mentioned above. The frequencies of those three peaks are around 0.066, 0.123 and 0.196; their corresponding periods are $1/0.066 \approx 15\ days$, $1/0.123 \approx 8\ days$ and $1/0.196 \approx 5\ days$. Therefore, it seems that, except the yearly cycle, the PM2.5 concentration in Beijing also has approximated half-month and weekly cycles. The residents in Beijing can make use of this result to take some preventive actions in advance.

6. Conclusion and Next Steps

In this project, ARIMA(1,0,1) has been verified as a good model to fit our analyzing data in terms of both complexity and robustness. It looks like there exists a positive relationship between the concentration of PM2.5 in today and tomorrow. In addition, we also identify the approximated yearly, half-monthly as well as weekly cycles of PM2.5 concentration, which might be helpful for the residents in Beijing to prevent the air pollution ahead of time.

For the future work, I may consider to implement Deep Learning models to analyze the data. One of the biggest advantages of Deep Learning is that people just need to feed the data into the model without many subjective judgements. Right now, more and more companies start to use RNN(LSTM) model to predict time series data. For example, Uber AI Lab sucessfully used LSTM to forecast "Engineering Extreme Event at Uber" and their result of prediction is shown

in Figure15[4]. We can see the prediction based on LSTM model seems very accurate, which is amazing. However, there are also some disadvantages of Deep Learning models. For instance, they are hard to interpret and also need lots of data to train. In addition to trying Deep Learning models, we also want to find out the relationships between PM2.5 concentration and other variables, such as wind power, temperature and humidity, so that we may improve our analysis to help reduce the air pollution more efficiently in Beijing.

7. Tables and Figures

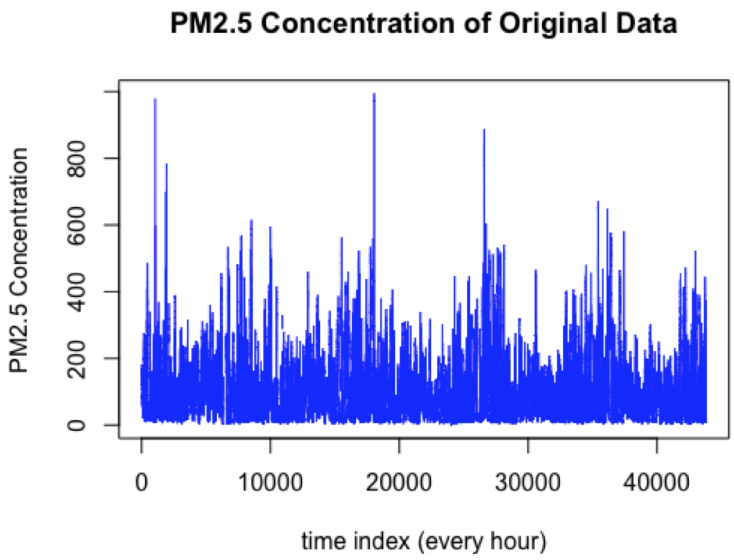

Figure 1: The PM2.5 Concentration of Oringinal Dataset

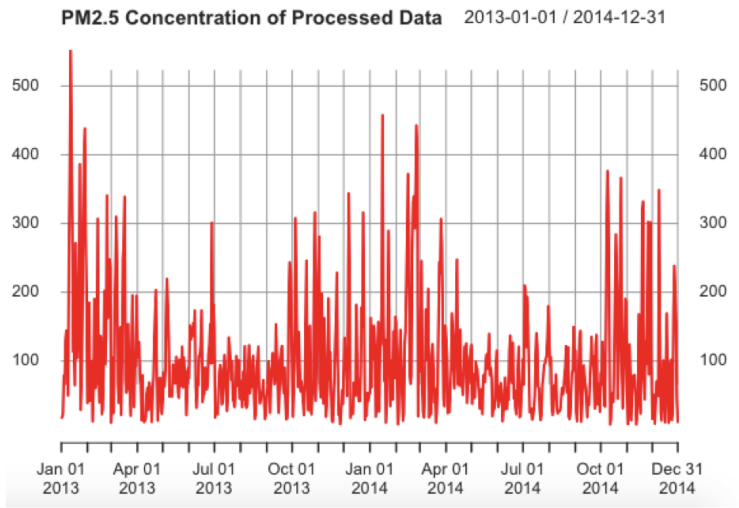

Figure 2: The PM2.5 Concentration of Preprocessed Daily Dataset

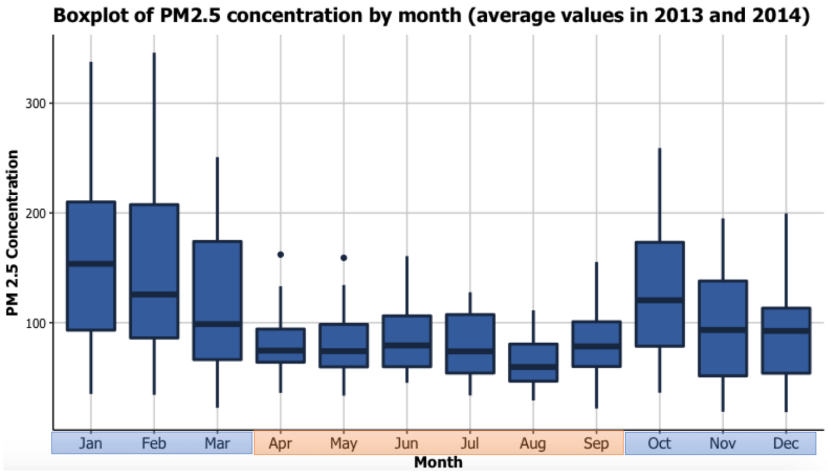

Figure 3: The Boxplot of PM2.5 by Months

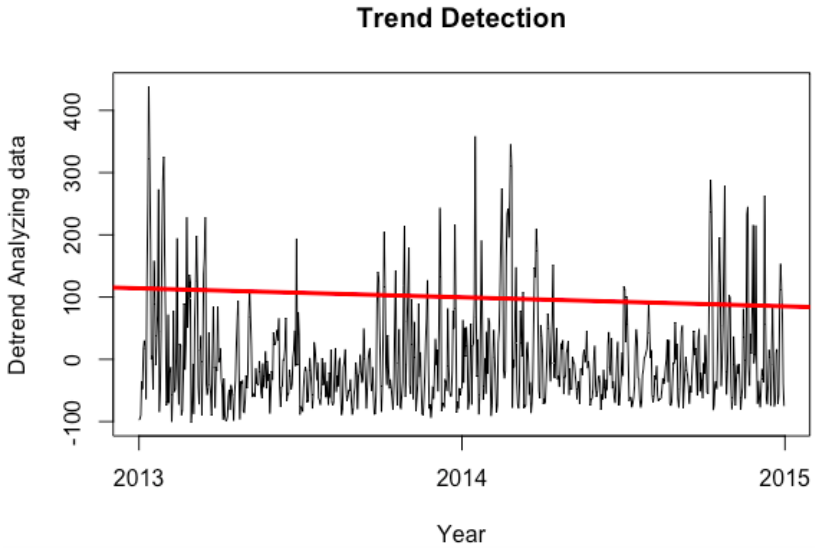

Figure 4: The Trend Detection of Analyzing Dataset

```
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
```

Figure 5: The Summary of Trend Detection

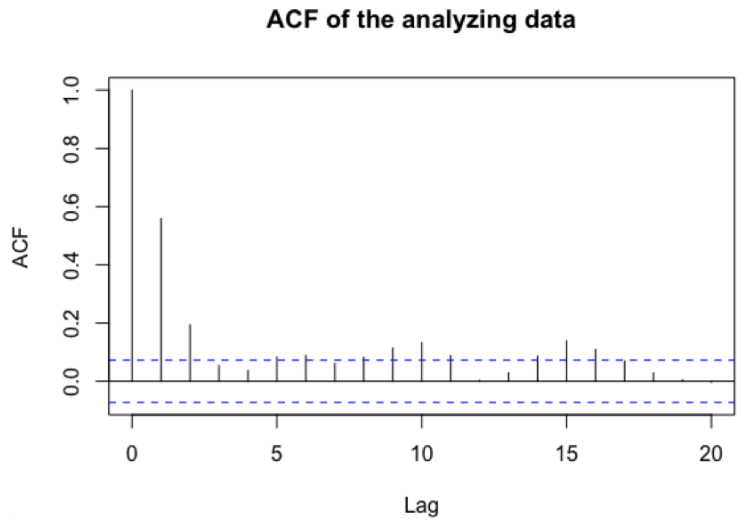

Figure 6: The ACF of Analyzing Data

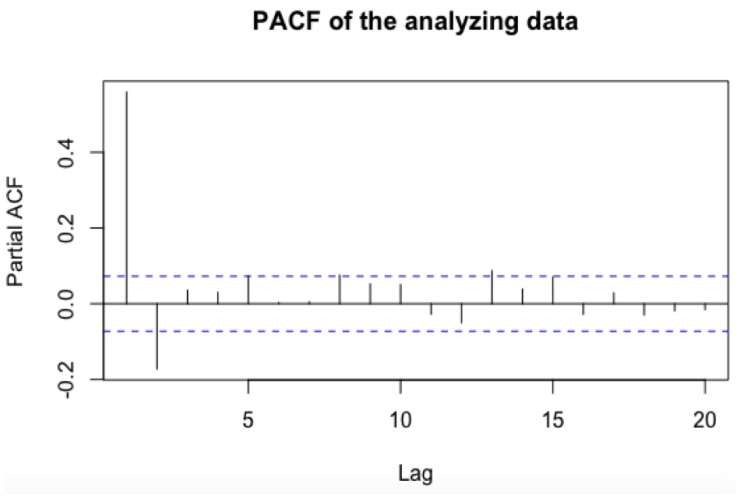

Figure 7: The PACF of Analyzing Data

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

Figure 8: Fitted ARIMA Models Based on AIC and BIC

```
$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
```

Figure 9: Estimated ARIMA(1,0,1) model

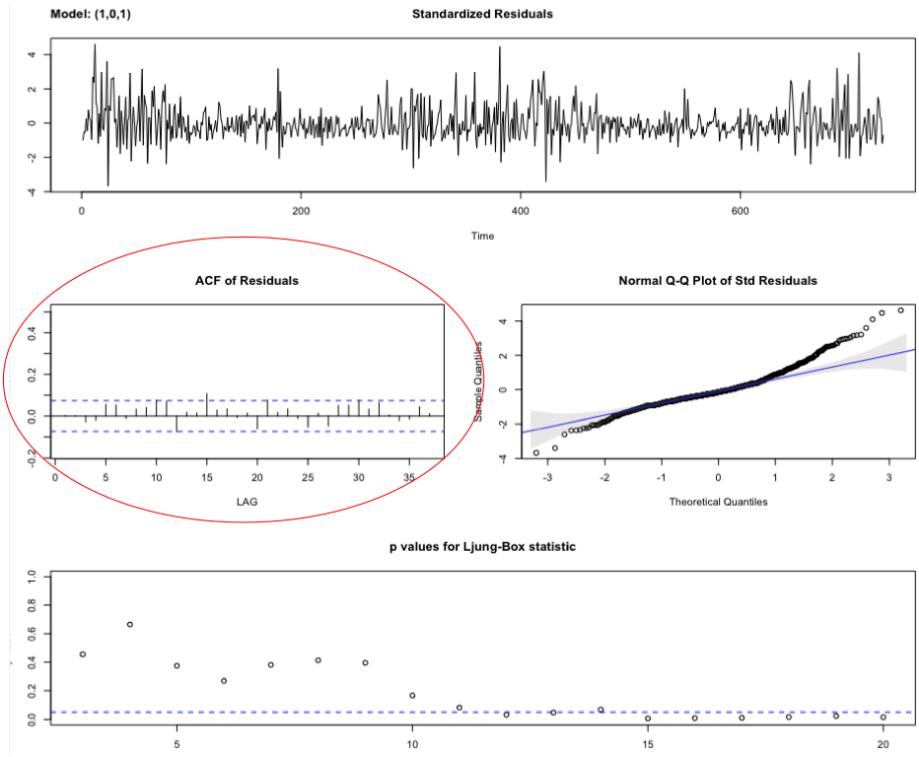

Figure 10: Prediction from ARIMA(1,0,1) model

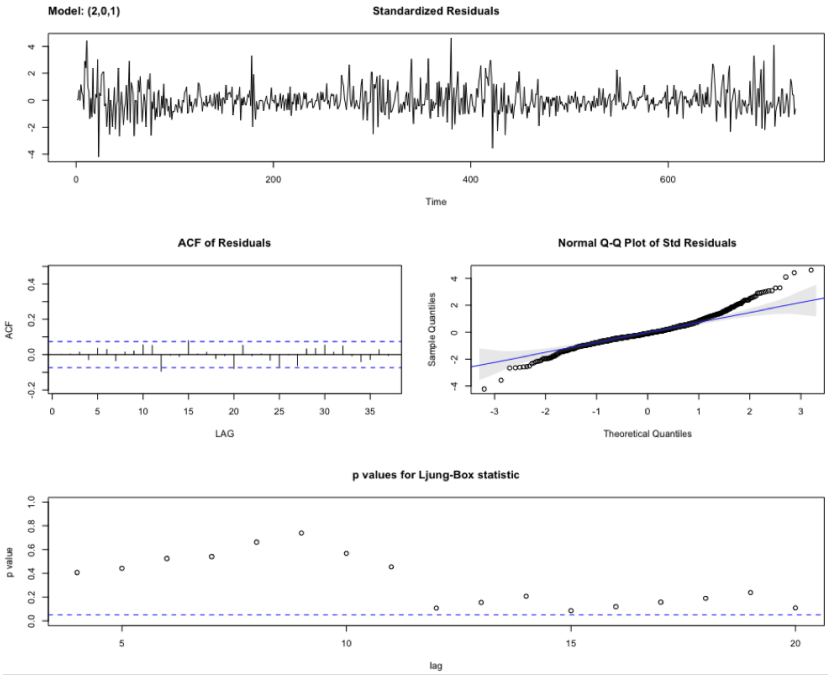

Figure 11: Model Diagnostics of ARIMA(1,0,1) model

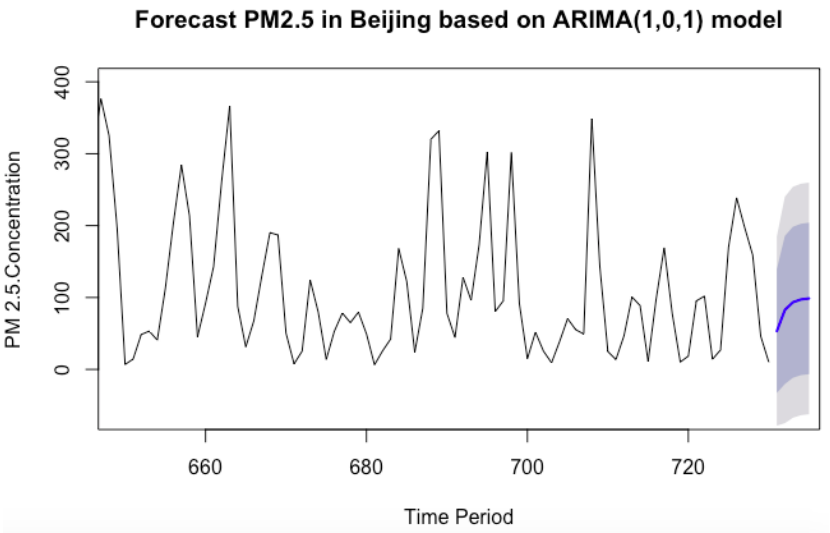

Figure 12: Model Diagnostics of Detrended Dataset ARIMA(2,0,1) model

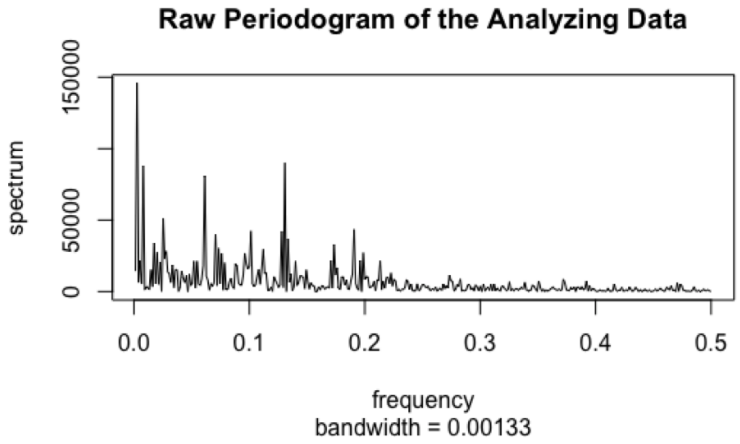

Figure 13: Raw Periodogram of the Analyzing Data

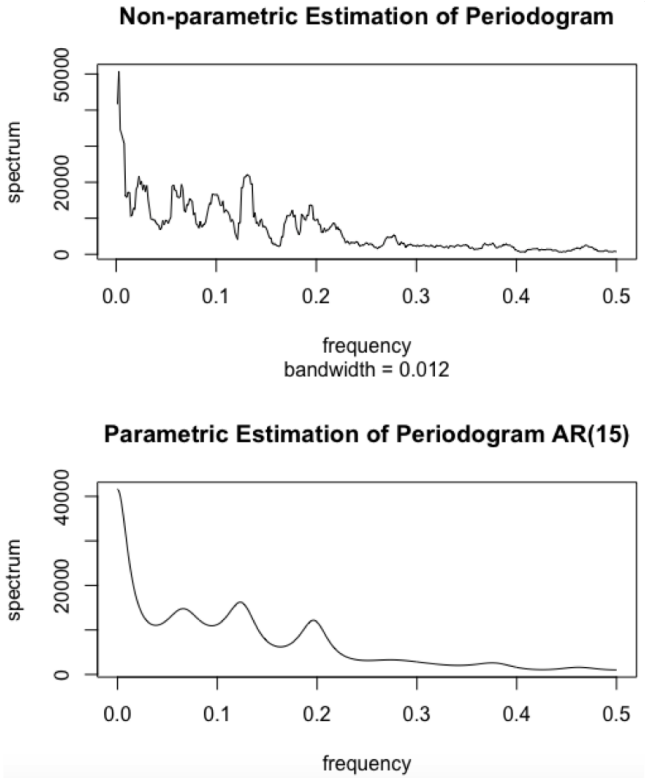

Figure 14: Non-Parametric and Parametric Estimation of Periodogram

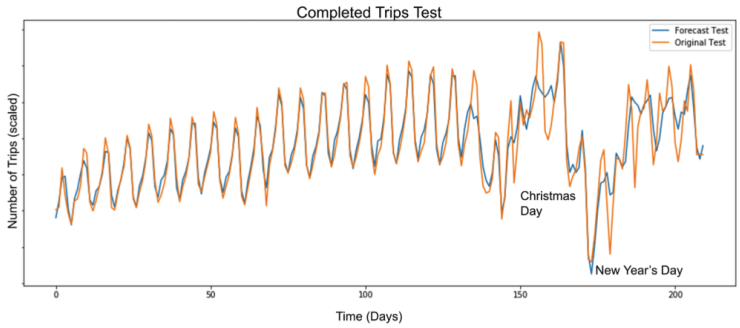

Figure 15: LSTM Result from Uber AI Lab

AQI Category	Index Values	Previous Breakpoints ($\mu g/m^3$, 24-hour average)	Revised Breakpoints ($\mu g/m^3$, 24-hour average)
Good	0 - 50	0.0 - 15.0	0.0 – 12.0
Moderate	51 - 100	>15.0 - 40	12.1 – 35.4
Unhealthy for Sensitive Groups	101 – 150	>40 – 65	35.5 – 55.4
Unhealthy	151 – 200	> 65 – 150	55.5 – 150.4
Very Unhealthy	201 – 300	> 150 – 250	150.5 – 250.4
Hazardous	301 – 400	> 250 – 350	250.5 – 350.4
Hazardous	401 – 500	> 350 – 500	350.5 – 500

Figure 16: Revised PM2.5 AQI Conversion Scale from Sep 09, 2013

8. References

[1] Haynie,D.,(2017).The Clear Thing About China's Smog [online] Available at ⟨ https://www.usnews.com/news/best-countries/articles/2017-01-13/the-health-effects-of-beijings-smog⟩(15/03/2018 10:07)

[2] Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.[online] Available at ⟨https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data⟩(15/03/2018 10:12)

[3] Revised PM2.5 AQI breakpoints,(2013).[online] Available at ⟨http://aqicn.org/faq/2013-09-09/revised-pm25-aqi-breakpoints⟩(15/03/2018 10:19)

[4] Laptev, N., Smyl, S., Shanmugam, S.(2017). Engineering Extreme Event Forecasting at Uber with Recurrent Neural Networks. [online] Available at ⟨https://eng.uber.com/neural-networks⟩(15/03/2018 10:25)