A First Look at the Chinese PM2.5 Data —
A Simple Prediction and Time Series Analysis in Big Cities

Group 4: **Chen Li**
Kejia Shi
Liqiang Pu
Yanan Huo

Abstract

The report makes predictions of air quality at selected time periods based on the air quality index(AQI) within the day, and points out the tendency of the air pollution in China on seasonal level. Also, detailed pattern analysis and cross-city analysis are provided.

Introduction

This report studies the air pollution in China based on the air quality index (AQI) on the US State Air Net website. The data source we use is the historical data set of Beijing, Shanghai and Chengdu from 2012 to 2015. We built a model to predict if it is healthy to go outside after dinner. We are also interested in the differences between weekdays and weekends. Based on the findings of seasonal trends and differences across cities, we perform a time series decomposition and use models and plots to quantify and visualize them. The government policy analysis for randomness part is provided at last.

Stateair.net China daily AQI datasets, the data source provides historical data for big cities from different years.

1 Data Cleaning

We combine the four-year dataset (2012-2015) for Beijing, Shanghai and Chengdu into a big dataset, which contains 35064 rows and 767 negative value of AQI within these rows.

We find 759 negative 999's and 8 small negative values. It seems meaningless to include negative values. Nevertheless, the small negative values, such as -1, -2, -15, is resulted from the lots of material from monitor's filters to the air. The large negative values, such as -999, are classified clearly to be due to the instrument malfunction or invalid. Thus, treating the large value as missing data is somehow necessary and we set them as NA's. Note missing data are not common among all data and nearly all of them comes fill in whole days, which means they would not constitute a big problem if we ignore them and calculate some characteristic values such as maxima and standard derivations of days.

In this explicit data set, we pick up the AQI for every hour including missing value as columns, that is the variables Hour0 to Hour23 in the Table 2. We calculate mean value, maximum and standard deviation of index within a day, that is from 0 am to 23 pm. Besides, from the criteria of the US State Air Net website, we also define our mean value and max value into this seven groups. The detail of this classification is in the Table 1.

Table 1. Classification criteria

AQI	0-50	51-100	101-150	150-200	201-300	301-500	>500
Group	Good	Moderate	Unhealthy for sensitive	Unhealthy	Very Unhealthy	Hazardous	Beyond Index
number	0	1	2	3	4	5	6

Table 1 describes the detailed classification of the Air Quality Index from the US State Air Net. According to the classification, we define the corresponding number as levels of the factor in our data set.

We create a new variable "week", including two factors "weekend" and "weekday", to divide the Beijing datasets into two groups. According to the calendar, we assign 2012/1/1 as weekend (Sunday), and the next seven days is a combination of 5 weekdays and 2 weekends, this pattern will persist until there's only remain 4 days (assign as weekdays). After that, we change some days' values according to the holiday arrangements.

Table 2. The rudimental data set

Year	Month	Day	Hour0	…	Hour23	mean	max	sd	week
2012	1	1	303	…	114	50.27	303	75.74	weekend
2012	1	2	97	…	14	78.30	169	58.18	weekend
2012	1	3	16	…	14	14.80	24	3.93	weekend
…	…	…	…	…	….	…	…	…	
2015	12	29			470	331.875	556	115.71	weekday
2015	12	30	536	…	26	101.75	536	164.60	weekday
2015	12	31	28	…	235	70.875	235	68.94	weekday

Table 2 shows the rudimental data set we yield, while in this data set, the mean, max and standard value is based on the AQI of 0 am to 23 pm within a day, which is corrected then.

2 Weekday or Weekend?

The first question comes to our minds is whether the weekday and the weekday make any differences to the air pollution. We estimate there to be some tendency of weekday and weekend through this four year, since the transportation presents as an important part in the constitution of index.

As we know, the lifestyles of people in these big cities differ, for example, the traffic flow in weekdays may be more crowded than that in weekends, and that may have some effects on the air quality. In this part, one important key question is to find out whether different lifestyles between weekdays and weekends (include holidays) have influence on the air quality. Also for different time periods, the influences can be varied. Therefore, the other key problem is to look for the trends of the influences, and try to explain the deep causes of the trends.

As indicated from stateair.net, for indices from 101 to 300, the cautionary statement defines the air quality as unhealthy to different groups of people. Hence, we choose index>=100 to be heavy and index<100 to be light.

Using graphical analysis, giving a count plot to show the distribution of Heavy and Light day for each group, also giving some scatter plot to seek for the trends or patterns.

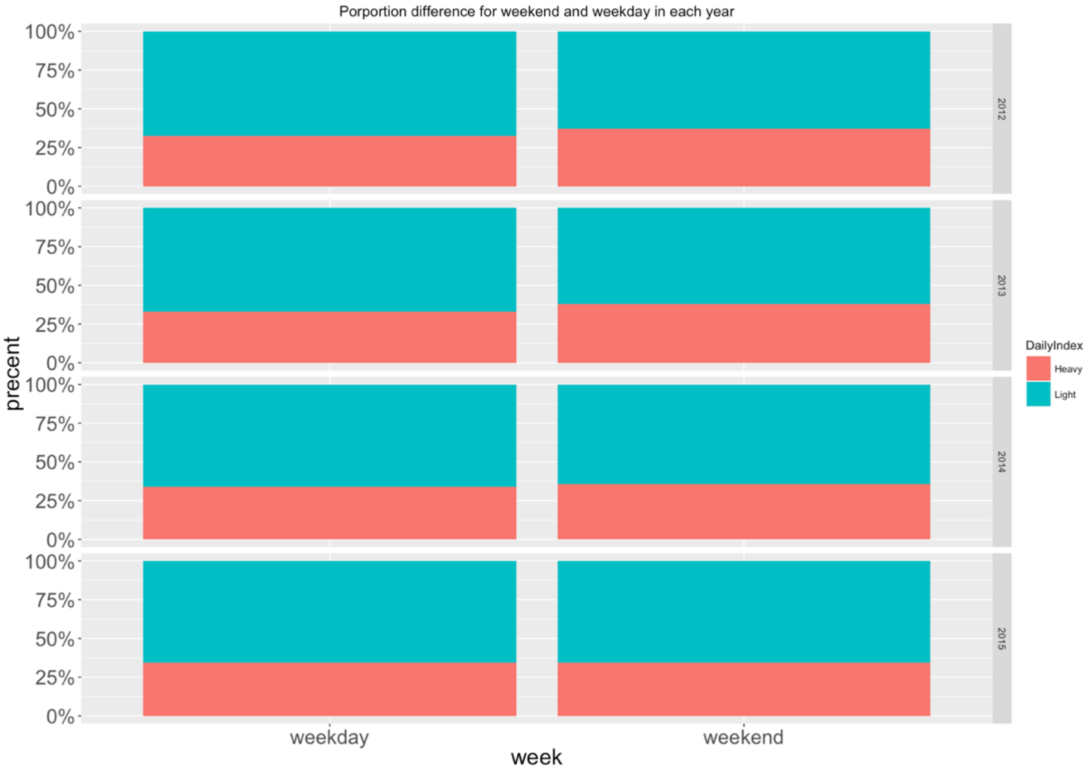


Figure 1. Weekday and Weekend Comparison

Table 3. Statistical Summary of weekday and weekend

Year	Week	Heavy_pollution_probability	mean_PM2.5_value	sd_PM2.5_value
12-15	weekday	0.334	92.4	91.0
12-15	weekend	0.363	95.0	91.2

Table 3. displays the heavy pollution proportion of weekday and weekend through the four years.

It seems that different lifestyles do have some influence on the air quality, according to the figure and table above. However, unexpectedly, the air quality in weekends are likely to be “heavy” with a highly average PM 2.5 values.

Of the next two pictures, left graph shows, from 2012 to 2015, there is a down trend for the probability of Heavy Day in weekends, but an up trend for that in weekdays. This suggests for the conclusion that weekends’ air pollution level may be lighter in the future. Right graph shows the patterns for months in one particular year, though there are some differences, the patterns are similar, which may due to the seasonal trend.

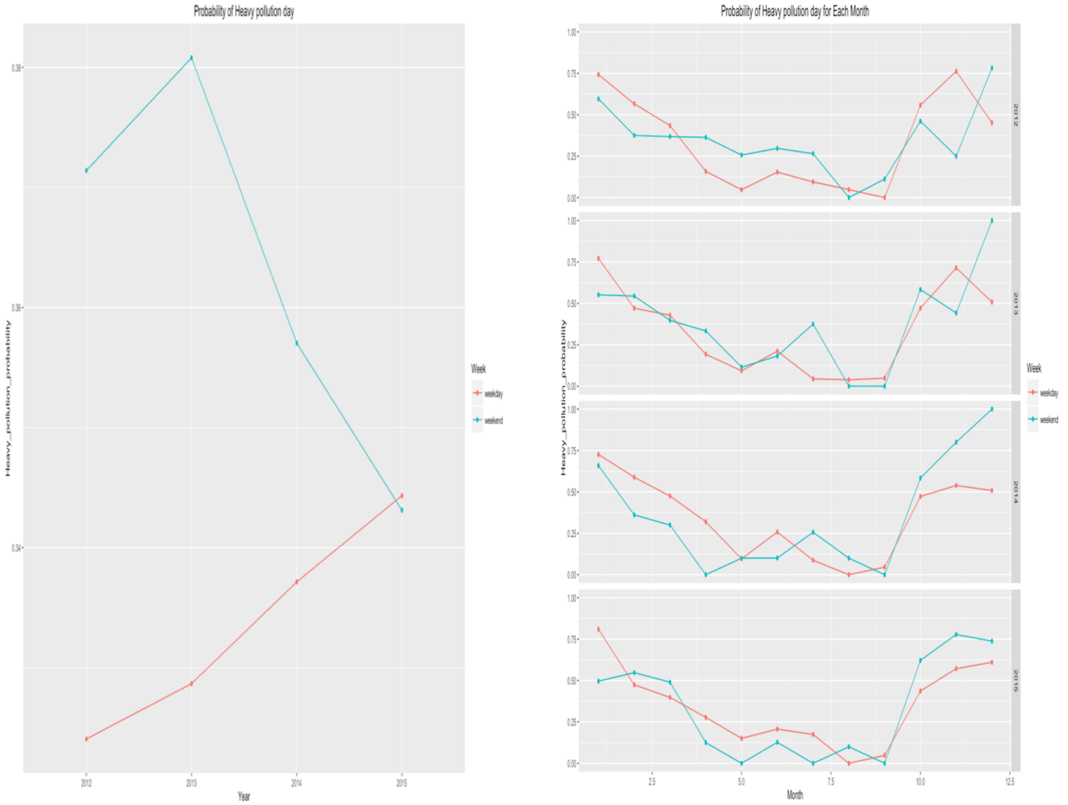


Figure 2. Probability of Heavy Pollution Day

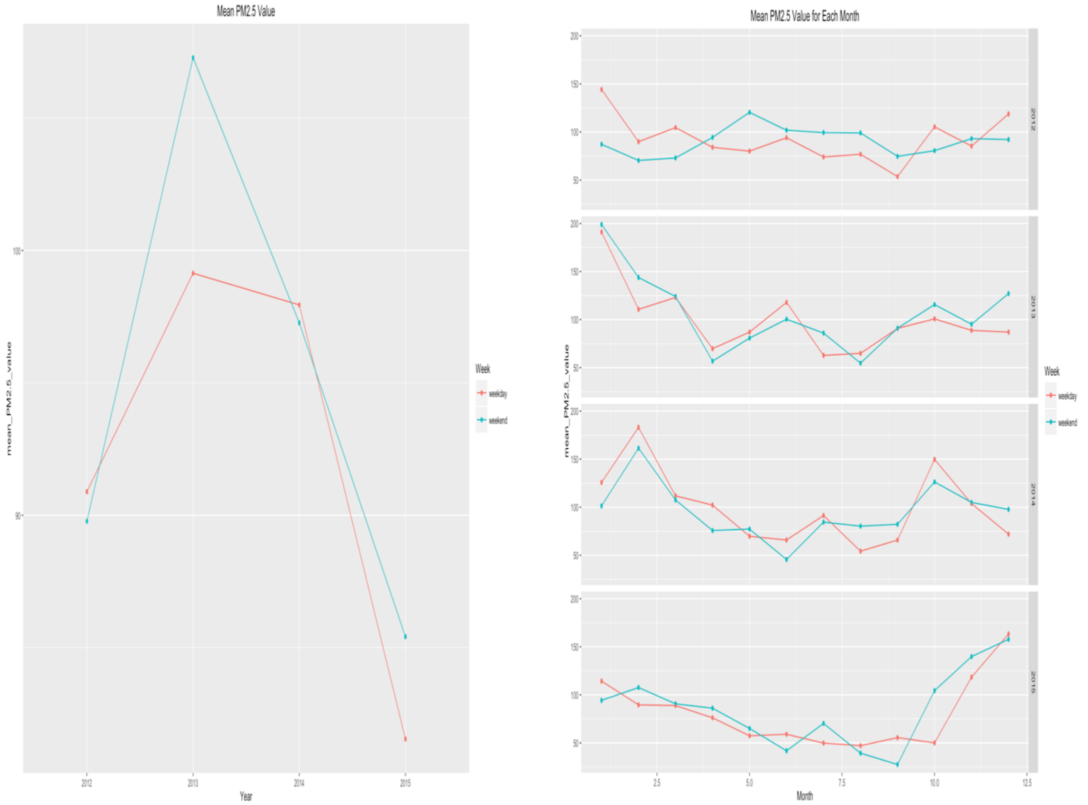


Figure 3. Mean PM2.5 Values

Not like the probability, both two groups reach the peak of average PM 2.5 at 2013, then down trend appears, which suggest that Beijing government may realize the pollution of air quality and have taken some actions after 2013. Also the patterns for months are alike.

3 Prediction

To further understand the pattern, especially for some important periods in a day, we try to build a model to predict the value of the average AQI from 8 pm to 10 pm, according to the previous AQI values within 24 hours and several relative variables.

After we pick up the 24-Hour variables, we recognized that we have made a mistake in calculating mean and max values. Because we treat the average AQI from 8 pm to 10 pm as dependent variable, the overall mean value should not involve these three hours, neither the max value or the standard deviation. Thus, we redefine our characteristic variables, using the AQI value of 11 pm of yesterday and also that of this day before 8 pm to make sure that we have no other information lost. Here, we have to throw away the data of first day other than 11 pm, that is, we loss 23 data points. Otherwise, we have to drop the 11 pm AQI of every day, where we loss 1461 data points. And the new data set is partly showed in Table 3.

Table 3. The final data set

Year	Month	Day	week	Mean21	Mean0	…	Mean 18	Max (index)	sd
2012	1	2	weekend	18.33	96.67		12.33	169(3)	51.18
2012	1	3	weekend	12.67	15.67		12.67	24(0)	3.93
2012	1	4	weekend	86.67	12.33		57.00	90(1)	28.79
…	…	…		…	…	…	…	…	…
2015	12	29	weekday	509.00	282.33		466.33	556(6)	115.71
2015	12	30	weekday	25.00	474.67		12.67	536(4)	164.60
2015	12	31	weekday	194.67	27.00		112.67	235(4)	68.93

Table 3 displays the final data set we use in building models. Meani's represent the average AQI of Hour i-1, Hour i and Hour i+1. Mean21 is the dependent variable and the others are explanatory variables, other then Year, Month and Day. We combined the mean value and the mean_index, max value and the max_index into two columns, while we call for them separately in the real data set.

Firstly, we selected linear regression to regress the whole variables, while the coefficient of mean is NA. The reason could be that the information of mean is overlapped by the information of the seven average variables. Thus, it is unnecessary to leave it in the model. Then, the full model is like following.

Mean21~
Month + mean0 + mean3 + mean6 + mean9 + mean12 + mean15 + mean18 + week + max + sd + max_index.

Then, we use the stepwise method to select the best subset of variables in the table, and we get the summary of the model and the diagnostic plots from R. The model we selected out has a 1368.136 mean squared error. Here we find out that the variable mean12 is insignificant by forth and back stepwise. By definition of stepwise process, we have enough reason to believe that if we drop off mean12, the AIC value of the model could increase. For the significance, we also examine the model without mean12, which yields that the mean squared error is 1370.552, which has little change of the stepwise-selected model. There is no big gap between whether we drop mean12 or not. While, to include mean21, it is somehow difficult to interpret the model, since its insignificancy.

Finally, the model we use to predict the average value of time 20 to time 22 uses following variables: month, mean0, mean9, mean15, mean18 and max. The R result is in figure 5.

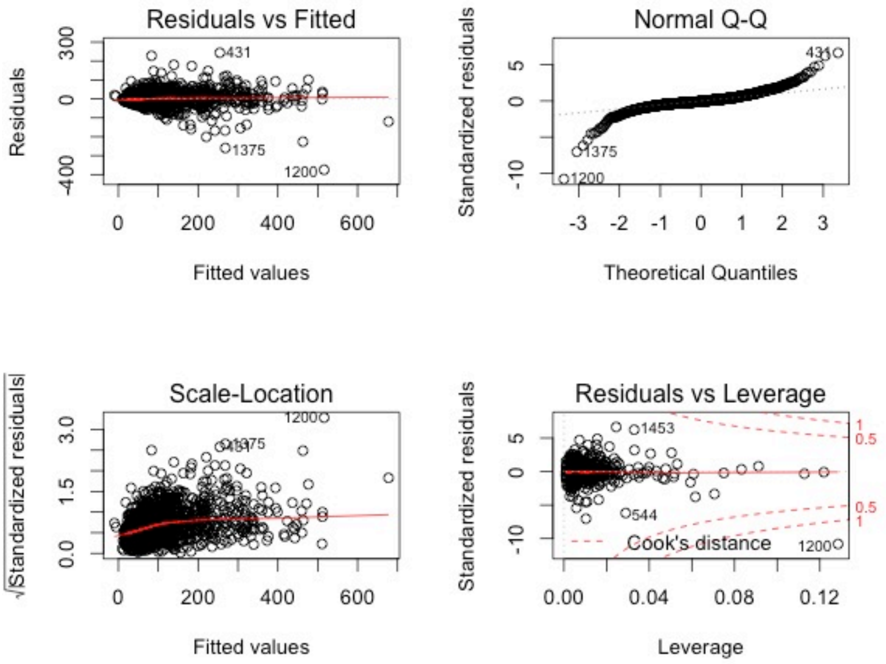
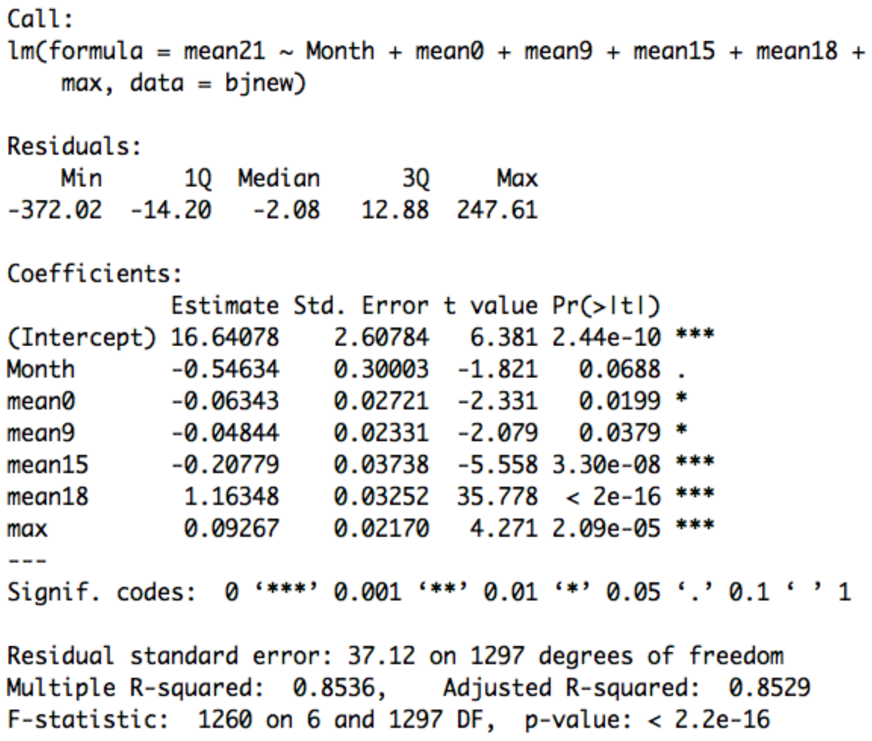


Figure 4. Diagnostic plots of the stepwise model



```
Call:
lm(formula = mean21 ~ Month + mean0 + mean9 + mean15 + mean18 +
    max, data = bjnew)

Residuals:
    Min      1Q  Median      3Q     Max
-372.02  -14.20   -2.08   12.88  247.61

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.64078    2.60784   6.381 2.44e-10 ***
Month       -0.54634    0.30003  -1.821   0.0688 .
mean0       -0.06343    0.02721  -2.331   0.0199 *
mean9       -0.04844    0.02331  -2.079   0.0379 *
mean15      -0.20779    0.03738  -5.558 3.30e-08 ***
mean18       1.16348    0.03252  35.778  < 2e-16 ***
max          0.09267    0.02170   4.271 2.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.12 on 1297 degrees of freedom
Multiple R-squared:  0.8536,    Adjusted R-squared:  0.8529
F-statistic:  1260 on 6 and 1297 DF,  p-value: < 2.2e-16
```

Figure 5. Summary of the final model

Certainly, we explore other regression methods, such as polynomial regression and regression tree, while among the methods we used, the mean squared error of prediction are all about 1400, which comes from the stepwise-linear regression. Once the linear model works invalidly, the random forest shall give a much lower mean squared error than linear model, but in this case, they are nearly the same. Thus, we do not have enough evidence to disapprove the linear model.

4 Time Series Analysis

Generally, the smog problem often gets worse in autumn and winter, especially in the North part of China for its centralized heating, apart from the other factors that happen all-year. However, to quantify the exact patterns of different seasons and to study the differences between cities, we have to visualize the pollution datasets.

4.1 PM2.5 Index Distributions

We plot the 2012-2015 PM2.5 index distributions by city in the following plot. Since the bad period of pollution is not in spring (and we are also missing the first 4 months' data in 2012 for Chengdu), we define the natural year starts from summer. The plot is shown in Figure 6.

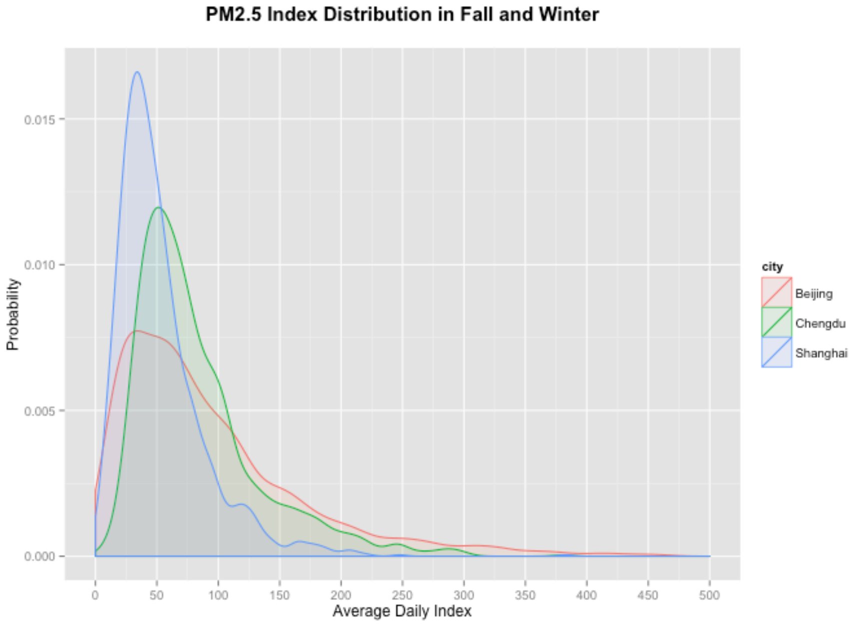


Figure 6. PM2.5 Index Distribution by City

As we can see from the plots, the curve of Beijing has a fatter tail compared to the other two cities, which demonstrate it having the worst air quality among the three cities we compared. Shanghai has the best air quality, with more than 80 percent of the days having indices less than 100. The data of Beijing has the largest variance. Then comes Chengdu and Shanghai having the smallest. We further plot the index distributions by season and natural year in different cities.

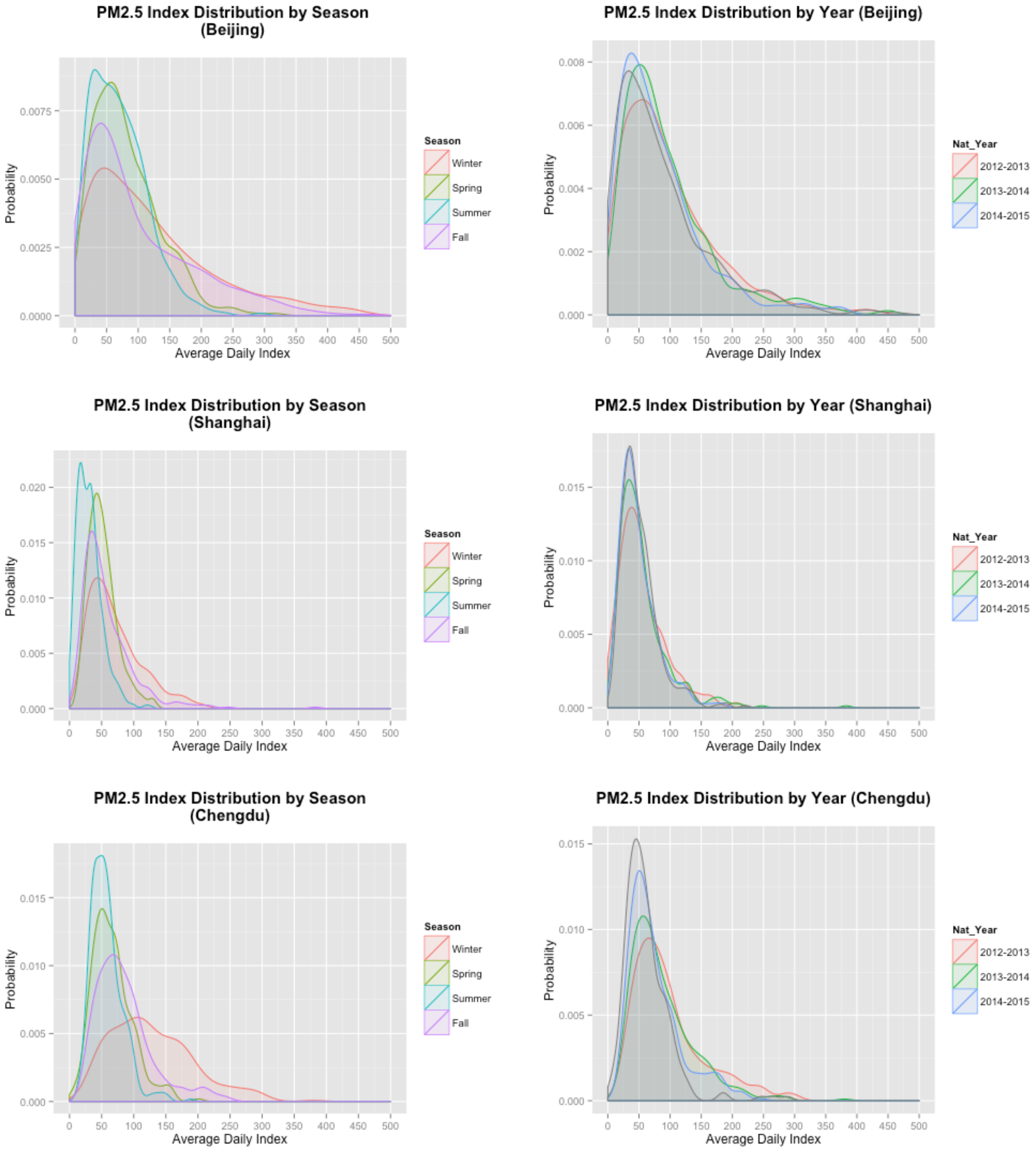


Figure 7. PM2.5 Index Distribution by Season and Natural Year

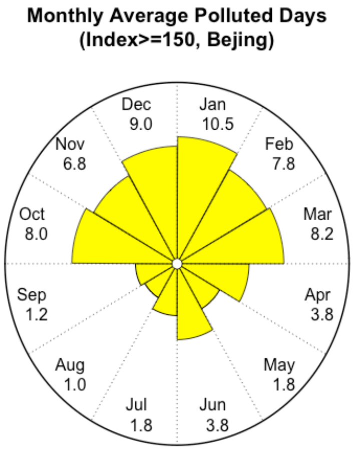
(Natural Year Starts from Summer)

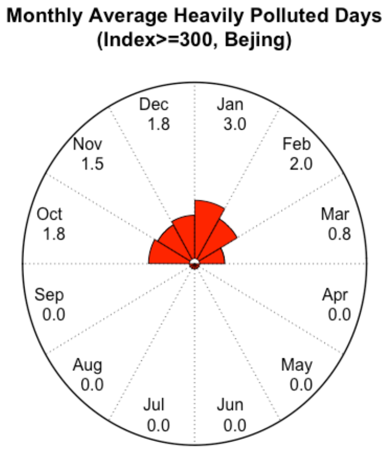
Air qualities in four seasons present the same rankings in the three cities, winter and fall being the worst. In Beijing, heavily polluted-day count in fall is very close to winter. In Chengdu, winter exceeds significantly in polluted-day count compared to the other seasons. We also notice from 2012 to 2015, in all three cities, less-than-100-index days constitute a larger and larger proportion in the yearly data. (Caution: the grey curves in the right side plots are not necessary needed, since it only reflects the rest of the data in our dataset.) To further cultivate citizen’s feelings about air qualities in daily life, we introduce dummies of polluted days and heavily polluted days.

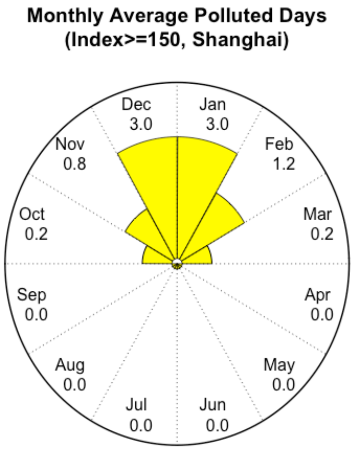
4.2 Monthly Average Polluted Days

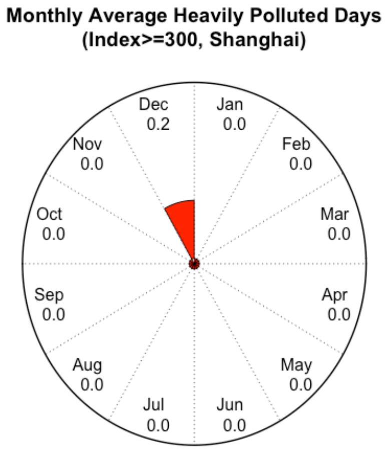
We add two more classification dummy variables, polluted and day count per month and heavily polluted and day count per month. As indicated from stateair.net, for indices from 101 to 300, the cautionary statement defines the air quality as unhealthy to different groups of people. Data above 300 is suggested that everyone should avoid all outdoor exertion. Hence, we choose index>=100 and index>=300 as two classification standards. Notice when the indices climb beyond 500, which is the worst case, we classify them into heavily polluted group. We also analyze the differences between the three cities, Beijing, Shanghai and Chengdu.

Now, with the two dummy variables, we plot the average polluted and day count per month and heavily polluted and day count per month as follows.


Monthly Average Polluted Days
(Index>=150, Bejing)


Monthly Average Heavily Polluted Days
(Index>=300, Bejing)


Monthly Average Polluted Days
(Index>=150, Shanghai)


Monthly Average Heavily Polluted Days
(Index>=300, Shanghai)

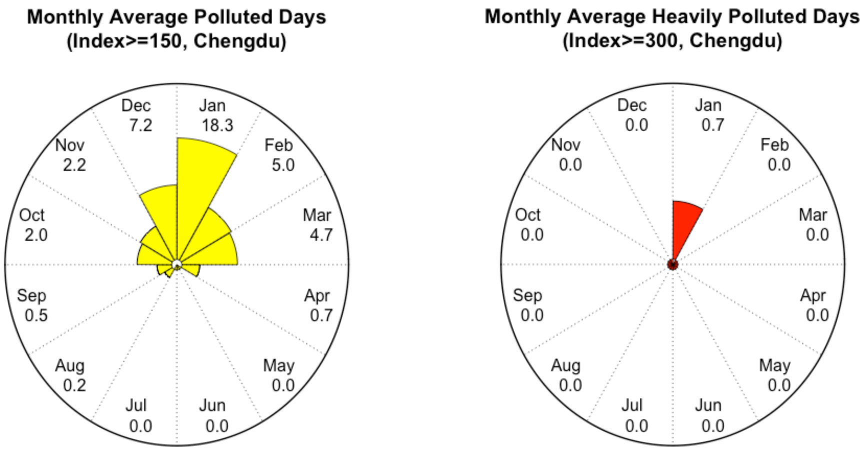


Figure 8. Circlar Plots of Polluted Days

It becomes easier to see that the (heavily) polluted days gather mostly from October to next March every year in Beijing, with the mean value being 8.38 days per month for the polluted days and 1.82 days per month for the heavily polluted days. Shanghai has a totally different pattern. The most polluted months are December and January, with average polluted days being 3 per month. Heavily polluted days only present in December. Chengdu has its most polluted months from December to next March. The highest count of polluted days in those months is January, 18.3, and the lowest is March, 4.7. Monthly heavily polluted situation only happens in January, average count being 0.7.

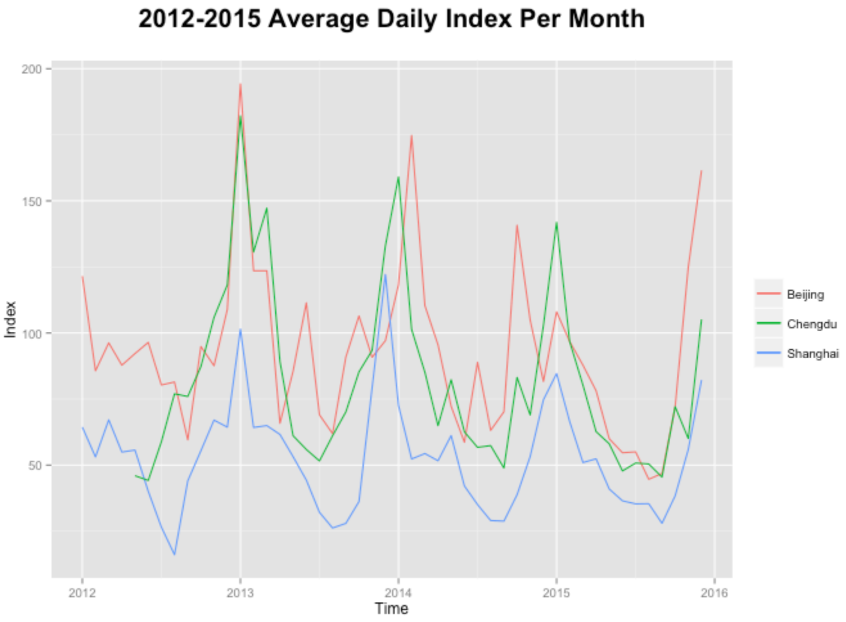


Figure 9. 2012-2015 Average Daily Index Per Month Comparisons

As we can see from the above picture of average daily index per month comparisons plot, we can also see that Chengdu has a continued decrease in its fall and winter index, which may due to a lot of reasons, such as pollution control policies or climate changes in specific years. The absolute index value is significantly lowered from about 180 to about 100. Shanghai has its yearly peak in 2013, which is December, 2013, being about 120. Beijing, though having the pattern of decreasing trend, still has the worst situation, making all the yearly peak values above 150. In 2015 specifically, the winter situation is worse than in 2014.

This coincide with our guesses. The northern cities may suffer from the smog problem worse than the southern cities. Also, different climates may cause different smog spreading patterns. Coast cities often have better air qualities and the inner areas often have the pollutants gathered for long. The January wind and precipitation patterns are usually very hard for pollutants to spread in Chengdu. Apart from these, we also notice that in Beijing, usually there are not completely with no polluted days in the other months, compared to the other two cities. This remains to be explored.

4.3 Seasonal Trend Decomposition

We decompose the three additive time series into trend, seasonal and random parts. A trend exists as a result of a long-term increase or decrease in the data. A seasonal pattern exists when a series is influenced by time-related. Here, it’s different months. The remainder part describes the randomness of the data, and the rises and falls are not of fixed period. We use the STL (Seasonal Trend Decomposition) method to decompose time series data and get the following graphs.

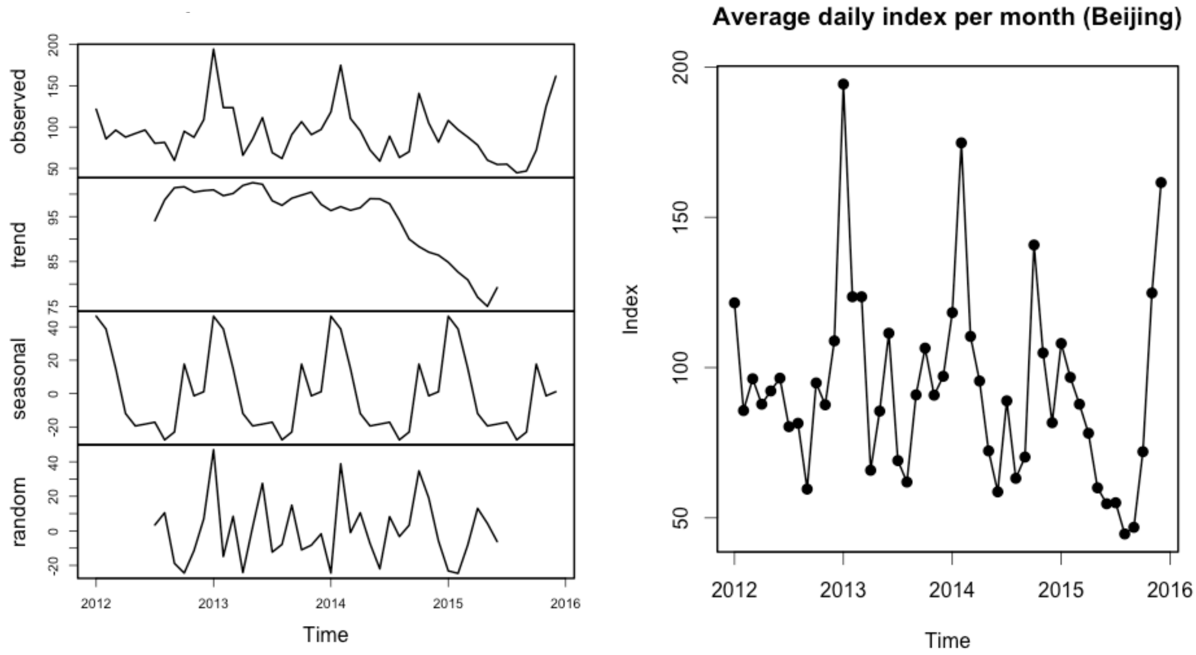


Figure 10-1. Trend Decomposition and Average Daily Index Plot in Beijing

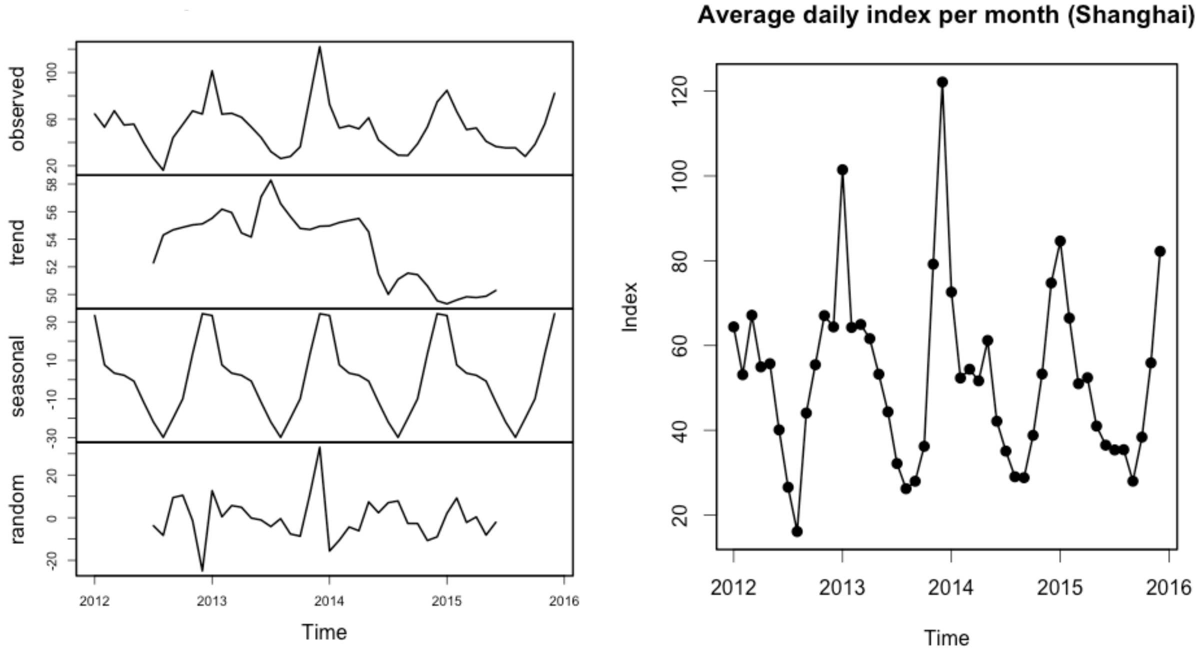


Figure 10-2. Trend Decomposition and Average Daily Index Plot in Shanghai

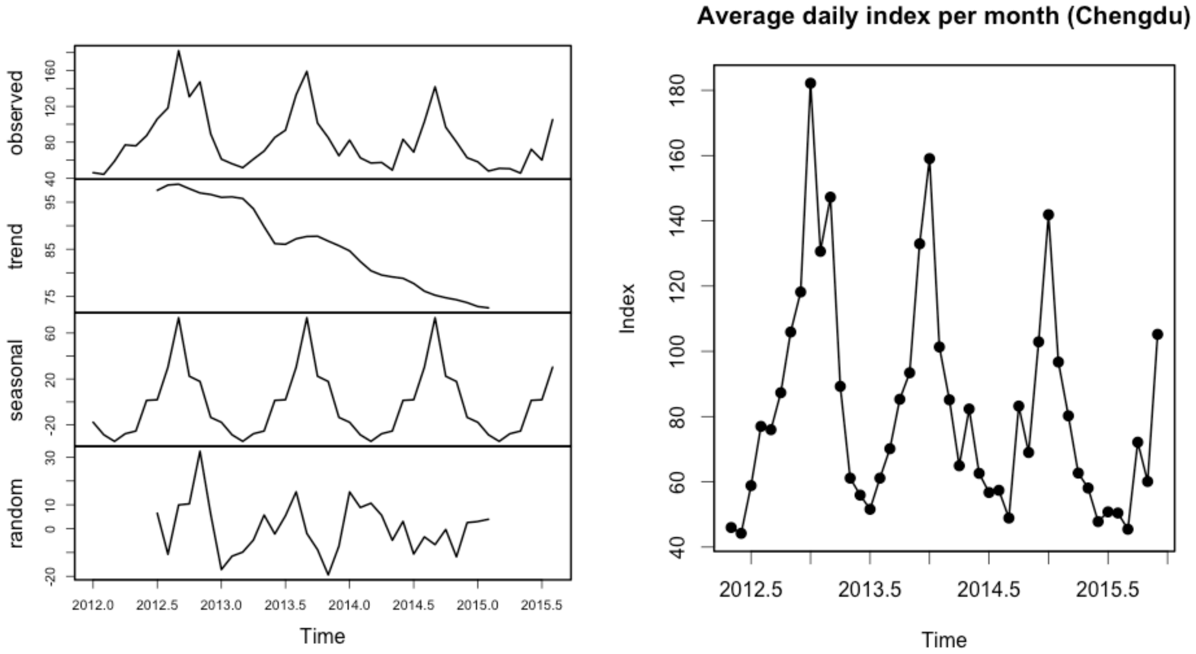


Figure 10-3. Trend Decomposition and Average Daily Index Plot in Chengdu

The above figures demonstrate the decomposition of monthly average index in the three cities. Seasonal composition in Beijing peaks in every January. The other little peak happens in October. Trend component shows that in 2014, the downward trend becomes more significant. Randomness satisfies our expectations. In Shanghai, the seasonal peaks in December and

January, and the trend also lowers significantly from mid 2014 after a slightly increase from the late 2013 to early 2014. In Chengdu, the peak of seasonal component lies in January and the trend is almost always downward.

Seasonal movements are usually fixed in different years, which makes the separated trends a cultivation of long term effect whether its government policies or climate changes.

4.3 Randomness Analysis – Special Policy Analysis

When there are important events, for example the Olympic Games, the air qualities need to be good. To ensure this, the government of China will take varies of actions, including policies and punitive measures. We want to capture the immediately effect of such policies, which constitutes the random part after decomposing the data.

We need to find the time intervals during which the air pollution get a great improvement and then search for the information related to these periods to find out whether it is caused by the government's policies or not. Though applied the approach to the data of three cities, in this report, we'll mainly talk about Beijing since it is the capital of China and it has the worst air quality among these cities.

In order to find the great improvement of air quality. we focus on the changes of the values of PM2.5 in a short time. we set every two days as a time interval and consider the mean of these two days' values as the interval value (for each day, since the standard deviation is large compared to its mean value, we prefer using the median). Then we calculate the difference of the adjacent interval value and pick out the minimal difference values (negative difference value means improvement of air quality) as our result and here's the plot.

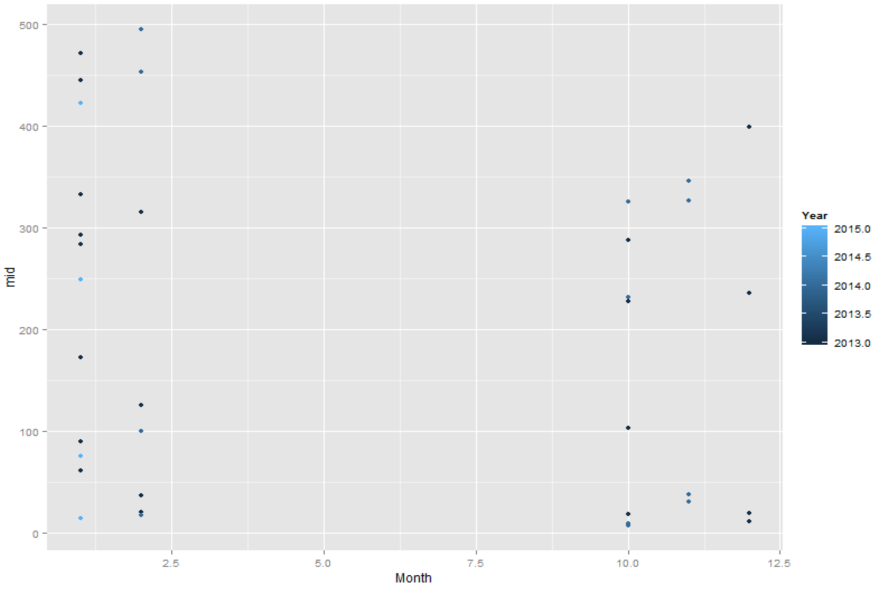


Figure 11. Removing Seasonal Effect (Beijing)

As we can see, the selected time intervals are all between October and March. It seems that the data still has some trends after removing the seasonal effects. Besides, we cannot find any great events related to these time intervals. These air quality improvements are more likely to be caused by the anomalous weather compared to government's policy. The result for Chengdu and Shanghai are provided below:

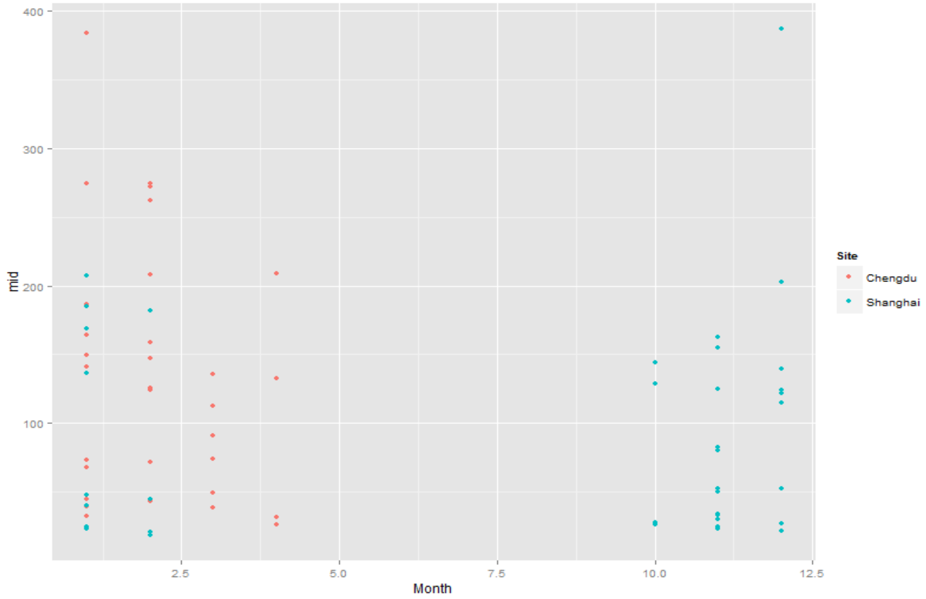


Figure 12. Removing Seasonal Effect (Shanghai and Chengdu)

We can see that the results of Beijing and Shanghai are similar. Their points locate in the same area while the points of Chengdu are between January and April. These maybe caused by the regional factors for Chengdu locates in a basin which makes it an anomaly. Besides, weather may also plays an important role.

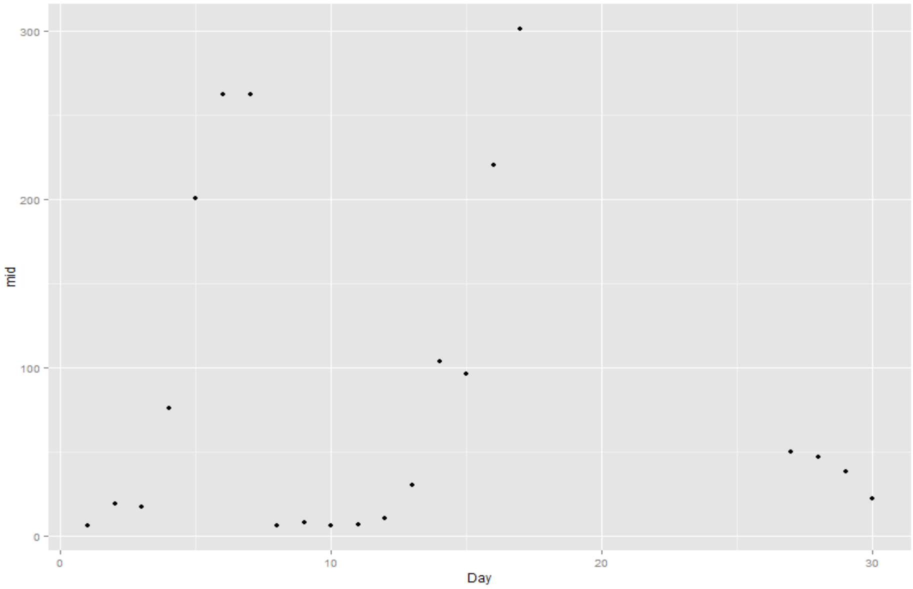


Figure 13. National Day, 2015 Scatter Plot (Beijing)

We expected to find some specific time intervals at the beginning of my plan during which some really important events happened and we think actions would be taken to ensure the good air quality. For example, the Oct 1 2015, which is the National Day of China and we had a great military parade in Beijing which means the air quality was supposed to be great during that interval. We plot the data of this interval (9.27-10.17).

We can see that the values of PM 2.5 experienced a decrease before Oct 1, but it is not a huge decrease in a short time and that's the reason why we didn't catch it by using our method. The improvement of air quality spent a relative long period. We look up the news and reports related to it and find that the government of Beijing took several actions such as traffic restriction about 10 days before the National Day. This led to a slowly but continuous improvement of air quality.

Besides 2015, we also analyze the data of National Day periods of 2013 and 2014. (We didn't have parade in these two year but some great celebrations were held during National Day).

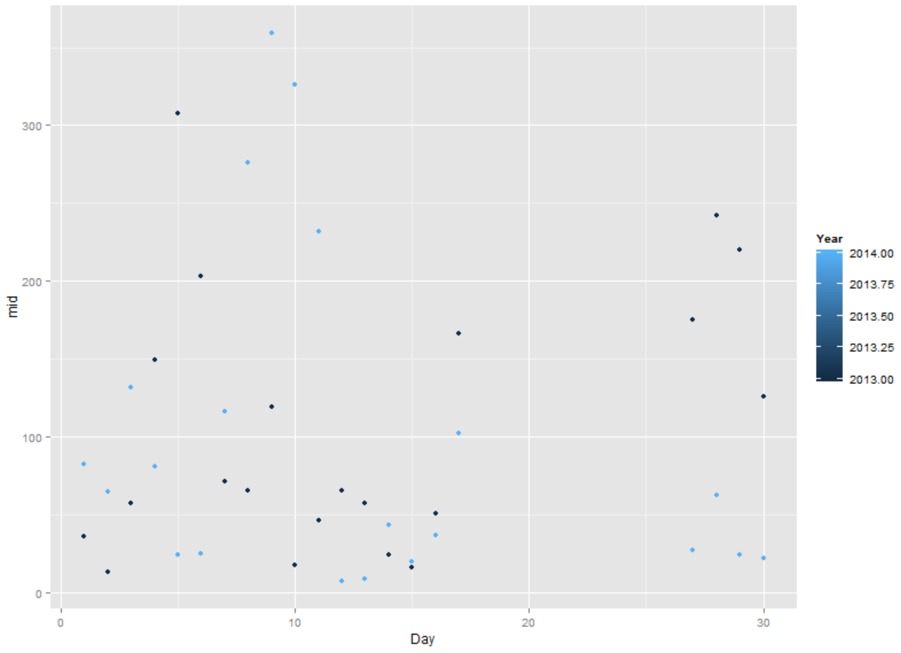


Figure 14. National Day, 2013&2014 Scatter Plot (Beijing)

We can see that in 2013, the air quality during the end of the September was not very good and it experienced an improvement in a short time. Compared to this, the air quality during the end of September in 2014 was already very good which is similar to 2015. The reason may be that the government changed their plan for improving the air quality in 2014 and it seems more feasible and effective so they continued using it in 2015.

As a conclusion, compared to solving the air pollution problem in a short time, I think the government of China prefer preparing for it in advance and spending a relative long time to achieve the goal. It is difficult to improve the air quality in a short time unless a strict traffic restriction or force all the large factories to close for several days and we know it is impossible to take actions like that. The government needs to find feasible ways to solve the problem without harming economy, for example, asking different factories to close at different time may be a better way.

5 Conclusion and Improvement

We captured the characteristics from the stateair.net dataset from many aspects. Generally, the exploration gives us a rather direct understanding of air quality data in prediction, composition and patterns in different cities.

It can't say for sure that the yearly trends for probability or daily average do exist since we only have 4 points (from 2012 to 2015). For future work, we may need add previous (year 2012 before) and future (year 2015 after) Beijing Air Quality datasets into our analysis, to see if the trends are persistent or the pattern would have some visible changes in future.

We could also replicate our prediction to more time periods. Moreover, if we have more historical data of every local monitor, then we can use the similar regression process to every nearby location and find the location that have the lowest PM2.5 value. In the end, it might become a good idea of a prediction app for daily use, which gives the public suggestions about good time to go out.

For the time series analysis, we have tried different models to cultivate the distributions and other patterns. Several useful time series models require data to be memoryless, which we could not apply. Due to the shortage of time, we could not take a dive into other models carefully. Also, we need to find related distributions to quantify what is plotted in the distributions by city, year or season.

As for the special policy studies, if we have more time, since we could not weigh the effect of such policies, we need to rule out weather and regional factors. (For example, it is more likely to rain in winter in southwest of China.) To get a better understanding of these problems, we are supposed to learn more about meteorology and geography.

References

State Air (U.S. Department of State Air Quality Monitoring Program). Retrieved on Feb 8, 2016 from http://www.stateair.net

Air Quality Index (AQI) Explained. Retrieved on Feb 22, 2016 from http://www.mass.gov/eea/docs/dep/air/aqi/aqi.htm

2015 阅兵期间北京限行细则. (2015 National Day Military Review Cars Restriction Policy). Retrieved on Feb 8, 2016 from http://news.tuxi.com.cn/news/184/1849360.html

2015 为迎接阅兵式北京已经进入“戒严”态势. (2015 Beijing’s Strict Policy Response of the National Day). Retrieved on Feb 8, 2016 from http://gw.yjbys.com/mingling/jieyanling/21406.html

China Holiday Information. Retrieved on Feb 10, 2016 from http://www.gov.cn/zwgk/2011-12/06/content_2012097.htm