Statistical analysis of air pollution data in Beijing, China

Bachelor's thesis

for acquiring the degree of Bachelor of Science (B.Sc.)

in Economics and Management

at the Ladislaus von Bortkiewicz Chair of Statistics
School of Business and Economics of
Humboldt-Universität zu Berlin

submitted by

Yang Chen

Student no.576421

First Examiner: Prof. Dr. Wolfgang Härdle

Second Examiner: Prof. Dr. Cathy Yi-Hsuan Chen

Thesis Supervisor: Dr. rer. nat. Sigbert Klinke

Berlin, August 20, 2018

Declaration of Authorship

I hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated resources. All passages, which are literally or in general matter taken out of publications or other resources, are marked as such.

Yang Chen

Berlin, August 20, 2018

Abstract

Air pollution is one of the most serious negative side effects in the process of industrialization. China, standing in the leading position of manufacturing and industrial production since the end of last century, has realized the importance of recognizing, identifying and reducing air pollution. This paper uses secondary data sets of suspended particulate matter(PM_5) collected from Beijing Municipal Environmental Monitoring Center(BJMEMC), to examine the patterns and identify the potential trend of air pollution in broadly Beijing area in the time period from 2013 to 2017 by doing descriptive and exploratory data analysis. Analysis shows that the great Beijing area is under severe particulate matters pollution but there is an obvious trend of decreasing shown in the data. Discontinuity test result shows no evidence consistent with a massive discontinuity at the cut-offs.

Keywords: Descriptive Data Analysis, Exploratory Data Analysis, Seasonal Analysis, Autocorrelation, Factor Analysis, Discontinuity Test, Time Series

Contents

1 Introduction

How to balance pollution and industrial growth is a big issue faced by many developing countries. Between 2005 and 2010, the number of deaths due to outdoor air pollution in China rose by about 5%[1]. In order to incentivize air quality improvement, china has been publishing a daily air pollution index (API) for major cities since 2000 and linking the API to local governmental performance evaluations, in which a day is defined as "blue sky day" when API is at or below 100. Since 2003, a city with at least 80% "blue sky days" in a calendar year (among other criteria) will be awarded as the "national environmental protection model city". This cutoff was increased to 85% in 2007(Chen et al., 2012).

The World Health Organization (WHO) and many other levels public health agencies have adopted fine particles that are smaller than 2.5 micrometers (PM2.5) or 10 micrometers (PM10) in terms of diameter as key metrics to control PM levels (Matus et al., 2012 cited from Holland et al., 1999), whose public health impact according to WHO(2006) is consistent in showing adverse health effects at exposures that are currently experienced by urban populations in both developed and developing countries. Particulate matters smaller than 2.5 μm originate primarily from combustion sources, for example coal and gasoline burning, while bigger particulate matters primarily produced by mechanical processes such as construction activities, road dust re-suspension and wind(WHO,2006).

A new ambient air quality standards *GB3095-2012* came into force nationwide on January 1, 2016, while Beijing, as one of the biggest air pollution sufferer and at the same time also as a pioneer of air quality protector in China has already

[1]The cost of air pollution: health impacts of road transport, Paris(2014):`http://www.oecd.org/env/the-cost-of-air-pollution-9789264210448-en.htm`(retrieved on 05.08.2018)

applied to this new standards three years earlier, on January 1, 2013.[2]In the new standards, the evaluation of suspended particulate matter $PM_{2.5}$ has been added to the measurements and the limiting values for 24-hours average value and one-year average value are also given. Great Beijing area are subsumed as category II(according to ambient air quality standards *GB3095-2012*[3] category I comprises mostly natural reserves and national parks; category II encompasses residential areas, industrial areas, rural areas and mixed areas), which has the limit value 35 $\mu g/m^3$ for one-year average and 75 $\mu g/m^3$ for 24-hour average, which is consistent with the interim target I(which is stated by WHO (2006) that, it is associated with about a 15% higher long-term mortality risk relative to the AQG(Air Quality Guidelines) level. In comparison, World Health Organization (WHO) has the guidelines of 10 $\mu g/m^3$ for one-year average and 25 $\mu g/m^3$ for 24-hour average[4].

In the year of 2013, the "Air Pollution Prevention And Control Action Plan" is issued by State Council on 10th September, 2013 (Document NO. GUOFA[2013]37)[5]. This plan includes optimizing industrial structure and reducing emission of multiple pollutants and so on. The year of 2017 is the end of the first stage of "Air Pollution Prevention And Control Action Plan"(2013-2017) and *People's Daily* has reported that in great Beijing area annual average of PM2.5 concentration has

[2]Ministry of Ecology and Environment of the People's Republic of China: `http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/dqhjzlbz/201203/t20120302_224165.htm`(retrieved on 05.08.2018)

[3]Ambient Air Quality Standards *GB3095-2012*: `http://210.72.1.216:8080/gzaqi/Document/gjzlbz.pdf`(retrieved on 05.08.2018)

[4]WHO air quality guidelines for particulate matter, ozone nitrogen dioxide and sulfur dioxide(2006): `http://apps.who.int/iris/bitstream/handle/10665/69477/WHO_SDE_PHE_OEH_06.02_eng.pdf;jsessionid=C7BB9291243F5ADF65AB2DCDCED28FD4?sequence=1`(retrieved on 05.08.2018)

[5]Air Pollution Prevention And Control Action Plan, translated by Clean Air Alliance of China(CAAC): `http://www.cleanairchina.org/product/6349.html`(retrieved on 05.08.2018)

reduced by 39.6% from 89.5$\mu g/m^3$ to 58$\mu g/m^3$ [6].

Despite all the perfection progresses on air quality achieved by China in media, the validity and trustworthy of air quality data that are published by the government is still remained to be questioned since multiple data manipulation and falsification cases were reported in different areas in China. In May 2018, Shanxi province, five people — including the former head of environmental protection in Linfen, Shanxi — were sentenced to prison of six months to two years for tampering with air quality monitoring equipment and falsifying data[7]. Andrews (2008a,b) first questioned the credibility of officially published data of Beijing and has brought this issue into public attention by presenting evidence that the API(Air Pollution Index[8]) has massive bunching below the cut-off together with inconsistencies between API values reported by the State Environmental Protection Agency (SEPA, `www.zhb.gov.cn`) and Beijing Environmental Protection Bureau (BJEPB,`www.bjepb.gov.cn`) at the cut-off. Ghanem and Zhang(2014) has expanded the regression discontinuity test to 113 cities during 2001-2010. Instead of using API(Air Pollution Index, they applied the McCrary-Test directly to the pollutants concentration data, which fulfill the requirement of continuity assumption of the McCrary (2008) test. Chen et al. (2012) proceeded formally an econometric analysis on the validity of the air pollution data. They brought evidence of anomalies around the cut-off based using the official data published by the government across 37 large cities in the time period from 2000 to 2009.

The remainder of the paper is organized as follows. Chapter two offers description of data source, objective and scope of the project and data quality together

[6]The State Council of the People's Republic of China: `http://www.gov.cn/hudong/2018-02/01/content_5262720.htm`(retrieved on 05.08.2018, translated by the author)

[7]The State Council of the People's Republic of China: `http://english.gov.cn/state_council/ministries/2018/06/25/content_281476197866592.htm`(retrieved on 05.08.2018)

[8]This index has been changed into AQI(Air Quality Index) since 2013

with data preparation for the following analysis. Chapter three will proceed descriptive data analysis, which firstly offers insight into pollutants level comparing with the critical value for both daily average and yearly average. Then in the part of box plot, information about distribution of $PM_{2.5}$ in each area is displayed. In the last chapter, exploratory analysis firstly study the seasonal pattern in the time series and secondly run factor analysis. Last but not least, discontinuity test will be applied to daily data in all observation stations in order to find any evidence of possible data manipulation.

2 Data Description and Preparation

2.1 Data Source

BJMEMC(Beijing Municipal Environmental Monitoring Center), founded in 1974, is one of the first professional environmental monitoring agency in China. The main function of the monitoring center is to be responsible for environmental quality monitoring of environmental factors such as atmosphere, water, noise, soil and ecology in the city area, monitoring of various pollution sources, and emergency monitoring of sudden pollution accidents[9].

BJMEMC provides atmospherical measurements including $PM_{2.5}$, PM_{10}, NO_2, CO, O_3 and AQI. In great Beijing area, there are in total 28 observation stations including 12 observation stations in urban area(station No.1-station No.12), 11 observation stations in rural area(station No.13- station No.23) and 5 observation stations in traffic intensive areas(station No.24- station No.28). According to the new ambient air quality standards *GB3095-2012*, fine particle matters $PM_{2.5}$ concentration values for each observation station are measured and published hourly by Beijing Municipal Environmental Monitoring Center since the end of 2013.

The U.S. Department of State Data provides $PM_{2.5}$ data, which is available from the Mission China air quality monitoring program[10]. The air quality data are measured at the U.S. Embassy[11] in Beijing since 2008. $PM_{2.5}$ concentration in U.S. embassy is measured hourly.

[9]Beijing Municipal Environmental Monitoring Center:`http://www.bjmemc.com.cn/jsps/jsp/zxgk/zxgk.jsp`(retrived on 15.08.2018, translated by the author)

[10]`http://www.stateair.net/web/historical/1/1.html`(retrieved on 05.08.2018)

[11]geographic coordinates of U.S. embassy is (39.95, 116.47) which is very close to observation station No.6 Nongzhanguan (39.94, 116.46)

2.2 Objective and Scope of the Project

The primary objectives of the study are:

- analyze fine particle concentration data in great Beijing area to identify patterns
- explore the possibility of discontinuity and anomalies in official data around cut-off

While PM2.5 is known to be a better predictor for PM-driven acute and chronic health effects than coarse mass (Schwartz et al., 1996)[12] and in order to analyze potential inconsistencies among different data sources, concentration value of $PM_{2.5}$ is chosen to be the prime scope of this paper.

Data of great Beijing area is the main scope of this paper because that firstly Beijing is one of the earliest city in China, which started to measure $PM_{2.5}$ concentration and publish all measurement officially, that provides the possibility to analyze the longest time period. For the analysis of time series variables, this gives more chances to find potential patterns and development along the time. For sake of completeness, variables in the following table is selected for the quantitative analysis hereafter:

[12]Is daily mortality associated specifically with fine particles?(1996): `https://www.tandfonline.com/doi/abs/10.1080/10473289.1996.10467528`(retrieved on 05.08.2018)

Table 1: List of Variables

Data Source	**BJMEMC**	**U.S. Embassy**
Variable Abbreviation	$PM_{2.5}$	$PM_{2.5}$
Number of Variables	28	1
Unit of Measurement	$\mu g/m^3$	$\mu g/m^3$
Data Type	hourly	hourly
Time Period	05/12/2013 - 30/06/2017	05/12/2013 - 30/06/2017

2.3 Data Quality

Hence that the interval of a valid measurement is [0, 500], all of the data points with a value falling outside of this interval or has a missing value are marked as "NA". Result shows that all variables have missing values. The underlying figure illustrates the percentage of "NA" cases out of total observations of each observation station in each year.

Missing values in data sets could lead to significant problems in statistical analysis. It is obvious that station number 9(Botanischer Garten Peking: urban area), 16(Tongzhou New Town: rural area), 19(Longquanzhen: rural area), 24(Qianmen Dajie: traffic intensive area), 27(South 3rd Ring Road: traffic intensive area), 28(East 4th Ring Road: traffic intensive area) have missing value cases way more than other observation station. Vertically compared, The first year(2013) of application of $PM_{2.5}$ concentration measurements and publication has substantially relatively more missing value cases, station number 9 and station number 19 have more than 30 % missing cases and there are other 3 stations(number 21, 25 and 27) have more than 10 % missing cases.

Although the reason for missing values is not cleared officially, this maybe results from the period of probation or the installation errors. Year 2017 has more missing value cases as well, which occur mostly in station number 9, 16, 27 and 28. Station number 9 and number 16 have more than 15% missing cases while station number 27 and number 28 have even more than 25 % missing cases.

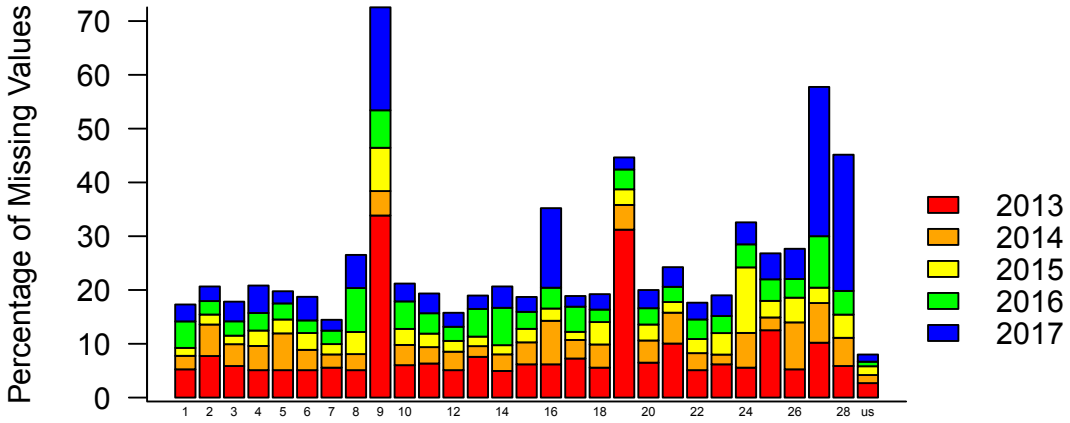

Figure 2.3.1: Missing Values in percentage of NA cases in total observations of each year, "us" refers to data collected by U.S. Embassy, see related quantlets: output/missing value

One possible treatment of missing values would be drop all the corresponding observations, which means loss of more than 30 % of all the observations. But then the reduction of sample sizes will lead to potential inaccuracy and inefficiency. Another approach to deal with missing values is mean-substitution. The good thing about mean substitution is that the mean value will be sustained but it will reduce the variance. The reason for choosing mean value is when there are

not too many outliers in the data sets. This method is not appropriate for this study because that $PM_{2.5}$ concentration has very obvious seasonal pattern, which will be illustrated in the following chapter. So missing values can not simply be substituted by the mean of all observations. A more advanced version of imputation would be conditional mean substitution. This method will calculate missing values based on the association with other variables. This would be meaningful if there is potential correlations among other pollutants variables. But in this study, $PM_{2.5}$ concentration is the only objective, so that conditional mean is also not a suitable solution for data incompleteness.

Variables in this study are time series and on hourly base, so that according to the two criteria(24-hour average and yearly average), data in the continuing 24 hours will be firstly grouped and calculated as daily average value. In this step, missing values will be dropped because $PM_{2.5}$ concentration will not change very rapidly in the next hour, so that the missing values will not play a big role in daily average calculation. In the second criterion of yearly mean, station number 9, 19, 27, 28 need to be treated carefully when compared vertically with other stations. Because high percentage of missing values in these four stations is worth questioning the validity and credibility of the measurements.

3 Descriptive Data Analysis

3.1 Daily Average

Firstly, daily averages of $PM_{2.5}$ in the whole period from 2013 to 2017 are grouped into three areas: urban area, rural area, traffic intensive area. Then the result are plotted together with the data collected from the U.S. Embassy, in order to compare with the critical value(75 $\mu g/m^3$, according to the new ambient air quality standards *GB3095-2012*, which is consistent with interim target I provided by World Health Organization(WHO)).

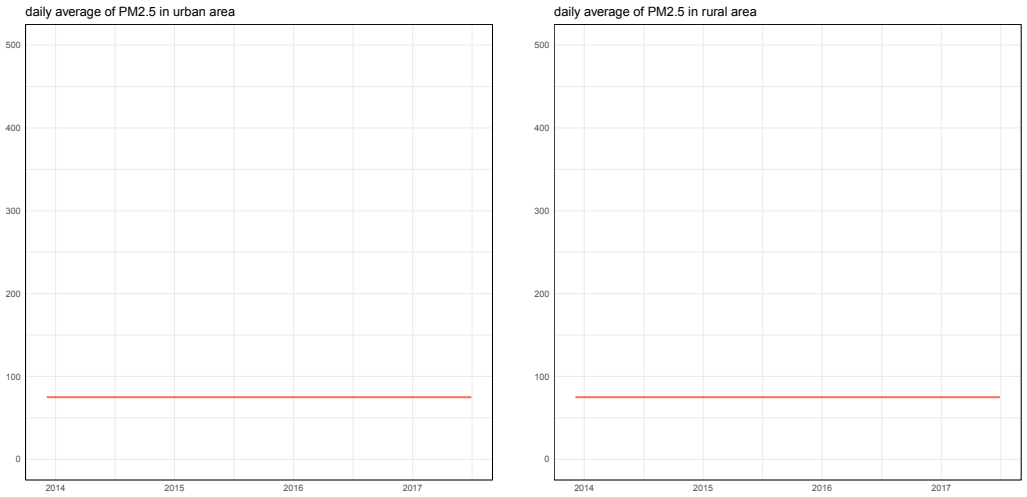

Figure 3.1.1: Daily Mean Values, left: daily mean values in urban area; right: daily mean values in rural area, see related quantlets: code/descriptive_analysis

The two graphs above both show four peaks during the whole period from 2013 to 2017. And the peaks appear round the change of the year(November, December, January and February). This means that particulate concentration level is much higher in winter time, compared to other seasons of the year. This seasonal pattern is shown very clearly in the graphs above. The left graph of averaged urban area shows that the peak in the winter of 2015(December 25th, 2015 at 495.2

$\mu g/m^3$) is the highest, which has almost reached the limit of measurement(500 $\mu g/m^3$), followed by 2016(about 450 $\mu g/m^3$ and 2013(400 $\mu g/m^3$). The year of 2014 has the lowest peak over the whole period(about 350 $\mu g/m^3$). The right graph shows the particulate concentration level in rural area. It can be seen that the peaks in the right graph is much lower than in urban area. But rural area has the the highest peak(December 1st, 2015 at about 420.6$\mu g/m^3$) at the same time as in urban area, which is the winter of 2015. The second highest peak of rural area is the winter of 2013, that has reached a value of 400 $\mu g/m^3$, followed by 2016 and the lowest peak has reached a value a little bit over 350 $\mu g/m^3$ in the winter of 2015.

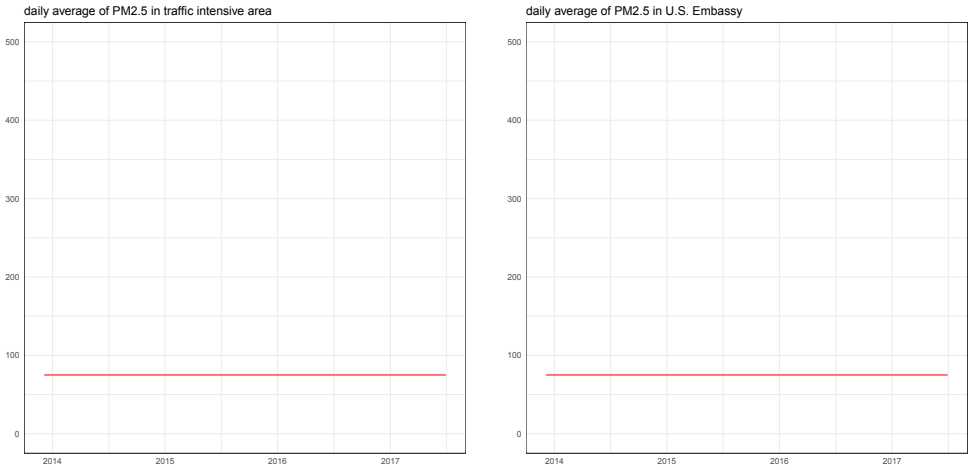

Figure 3.1.2: left: daily mean values in traffic intensive area; right: daily mean values in U.S. Embassy, see related quantlets: code/descriptive_analysis

Figure 3.1.2 shows the daily average value of traffic intensive area and the particulate concentration in the U.S. Embassy. Traffic intensive area has in general lower performance in particulate concentration level than U.S. Embassy. While the highest peaks in both graphs, which happened in the winter of 2015, have reached 470 $\mu g/m^3$. The second highest peak of traffic intensive area is in the winter of 2013, which has reached 400 $\mu g/m^3$, followed by the winter of 2014 with a peak below 400 $\mu g/m^3$ and the lowest peak of particulate concentration level is in the winter of 2016 at about 350 $\mu g/m^3$. U.S. Embassy has an extreme value in the winter of 2016(January 1st, 2017) at 454 $\mu g/m^3$, which is much higher than other observation in the same time period. And except for this outlier, the peak of winter 2016 is at the level of 400 $\mu g/m^3$. The third peak located in the winter of 2013 and has reached 420 $\mu g/m^3$. The lowest peak of particulate concentration is in the winter of 2014, which is below 400 $\mu g/m^3$.

Table 2: Percentage of days exceed critical value of 75 $\mu g/m^3$

	urban	rural	traffic	us
2013	0.407	0.444	0.296	0.37
2014	0.426	0.462	0.457	0.487
2015	0.41	0.424	0.415	0.443
2016	0.398	0.406	0.402	0.41
2017	0.382	0.387	0.389	0.393

Table 2: Percentage of days exceed critical value of 75 $\mu g/m^3$, see related quantlets:


exploratory_analysis

Table 2 shows the percentage of days in each year that exceed the critical value of 75 $\mu g/m^3$. Percentage value greater than 45% are marked as pink. Although in

all four graphs above, the highest peaks are in the end of 2015 and beginning of 2016 and the winter of 2014 has the lowest peak in most of the graphs. The year of 2014 in rural area, traffic intensive area and U.S. Embassy have the highest percentage of days, that exceed the critical value. This maybe can be explained that, in the years with higher peak, there are also more small values below the critical value. As it can be seen that in the second half of total period, extreme small values are more densely plotted. This variation in distribution may caused the different performance of each sub-period regarding different statistical applications.

3.2 Yearly Average

In this chapter yearly *mean* values of each station are firstly calculated in order to compare with the critical values of category II in the new ambient air quality standards *GB3095-2012*, which is consistent with interim target I provided by World Health Organization(WHO).

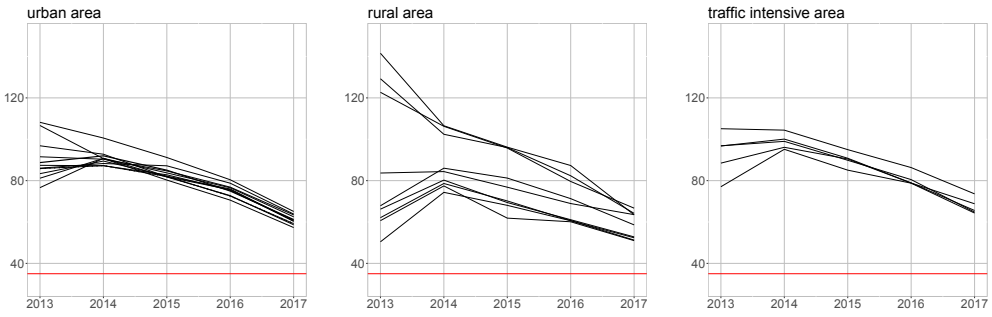

Figure 3.2: Yearly Mean Values, left: yearly mean values in urban area(dotted line is data collected by U.S. embassy, fat line indicates data from station No. 6 in comparison), middle: yearly average mean values in rural area, right: yearly mean values in traffic intensive area, see related quantlets: code/descriptive_analysis

It is very obvious that all three groups of observation stations have $PM_{2.5}$ concentration above the critical value of 35 $\mu g/m^3$. And most of the measurements fall into the interval of [80, 120]$\mu g/m^3$, which is two to three times of the critical value. There is no significant difference among all three areas except for rural area, which has measurements divided into two subgroups. This implies that great Beijing area is under serious pollution of particulate matters. And in all of this three graphs, there is a trend of downward in the development of the particulate concentration. To be noticed, data in the year of 2013 is only available in December and particulate matters are very highly associated with seasonal patterns, so the average value of year 2013 is not completely comparable with the following years. At the end of this five-year period, all three areas have reached a value under 80 $\mu g/m^3$.

The left graph above shows that all 12 observation stations have very similar behaviour. Except for the year 2013, varies in measurement values are mostly less than 15 $\mu g/m^3$. From 2014 to 2017 there is a holistic decrement in the yearly average value and it has reached downwards the level around 65 $\mu g/m^3$ by decreasing about 25 $\mu g/m^3$, which shows a great improvement in air quality control. To be noticed that the dotted line in the left graph indicates the measurements collected and published by the U.S. embassy and the fat line shows the particulate matters in station No.6, which is only in three kilometers distance with the U.S. embassy. The measurements in these two locations are very close to each other and the level of particulate matter in the U.S. embassy is lower than observation station No.6 with a difference of maximum 4.72 $\mu g/m^3$ and minimum 1.96 $\mu g/m^3$.

The middle graph shows the level of $PM_{2.5}$ concentration in rural areas. It can be seen very clearly that there are two subgroups in the graph, values measured in station No.13(Fangshan), No.14(Daxing) and No.15(Yizhuang) are much higher than other 6 observation stations in general but the differences decrease by the time flow and disappear at the year of 2017. By then, these two subgroups jointed. The

upper subgroup in general has higher value than observations in urban area and traffic intensive area, while the lower subgroup has lower values than urban area and traffic intensive area.

The right graph shows that in traffic intensive area, there is no significant big difference in particulate concentrations than other areas. But a decrease in the measurements can be observed and the value of particulate matters are slightly higher than urban area.

3.3 Box Plot

One possible way to study the distributional characteristics is enabled by box plots. For data description, box plots for urban area, rural area and traffic intensive area covering whole period from 2013 to 2017 are generated, in order to have a general look at the data sets and compare data performance along the time.

Data from all observation stations are firstly grouped as urban area, rural area and traffic intensive area. Then for each single day an average value for measurements in all stations in this group is calculated and served as *mean* performance for this group. It is shown in the box plots, that the distribution in all three areas have similar patterns. In the first year of 2013, there are only data limited in December and due to the size of data sets, there are no outliers. On the contrary, there are many outliers in other years and all outliers have extreme high values. The box plots for the year of 2013 is comparably taller than other years, which indicate that the distribution have a bigger range. While the box plots for the year 2014-2017 are relatively shorter. This implies that the distribution of the main middle part of all data points are quite similar.

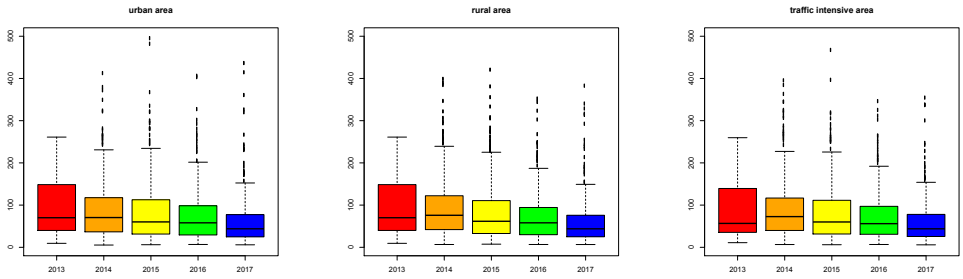

Figure 3.3: Box plot, left: box plot of pollutants in urban area, middle: box plot of pollutants in rural area, right: box plot of pollutants in traffic intensive area, see related quantlets: code/descriptive_analysis

From the left graph, a slight decrease in median value is observed, which is consistent with the yearly average values calculated in the last chapter. But the distribution in upper part and lower part is different. It is apparent that the distribution in upper part varies much more than the lower part. This phenomenon is relatively more significant in the year 2015 and 2017. In these two years, there are also more outliers at extreme high level, for example 495.24 $\mu g/m^3$ on 25th December, 2015 and 482.12 $\mu g/m^3$ on the first of December, 2015, compared to other years. The behaviour of particulate concentration in rural area is like the pattern in the urban area, but only with a slightly lower value. While in the right graph about data distribution in traffic intensive area shows a lower median value in year 2013 and higher level of extreme values at 467.43 $\mu g/m^3$, which is on the first of December, 2015.

4 Exploratory Data Analysis

4.1 Seasonal Analysis

Particulate matter concentration data from each area(urban area, rural area and traffic intensive area)are averaged in each group on a daily base. In order to apply to seasonal analysis, the time series are firstly calculated into monthly average. Then, these time series are applied to the autocorrelation function to compute estimates of the autocorrelation coefficients (ShumwayStoffer, 2011):

$$\rho(s,h) = \frac{\gamma(s,h)}{\sqrt{\gamma(s,s)\gamma(h,h)}} \tag{1}$$

and produce a plot showing autocorrelation coefficients together with confidence band in blue indicating if the autocorrelation coefficients are statistically significantly different from zero, which refer to a significant influence on the corresponding lags. By default, the confidence interval is 95%. In this study, degree of lags is set as 24. Because a two-year period should be enough to observe any seasonal patterns.

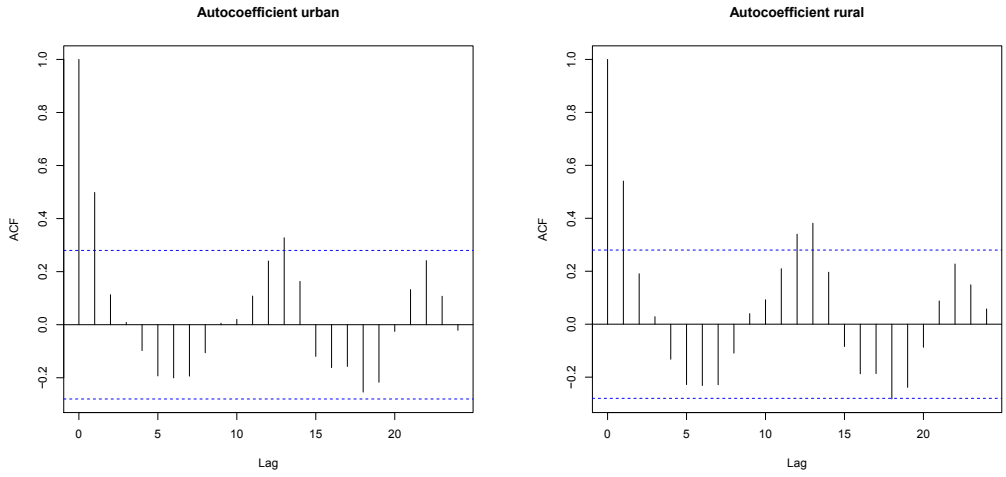

Figure 4.1.1:left:urban area; right:rural area, see related quantlets: exploratory_analysis

Among all areas, seasonal patterns in autocorrelation coefficients are shown in the graph above. The shape of the graphs is like a cosine curve with a period of 12 months. This repeated periodic pattern of 12 units is very clear in all of the four graphs, while the absolute value of the coefficients are mostly located in the not significant interval. In the first quarter, the first three lags are positive correlated and the absolute value of autocorrelation coefficients decrease. Then the following six coefficients turn into negative and the absolute value first increase till reach the lowest point at around the sixth lag, then the absolute value of coefficient decrease. In the last quarter of 12 units, autocorrelation coefficients run into positive again and the absolute value increase until about 0.3. This periodic pattern repeat in the next 12 unit only with relatively lower absolute value than the first 12 units. Observations in 12/24 months or one/two year apart are relatively strongly positively associated. While observations in 6/18 months or half/one and half year apart are relatively strongly negatively correlated. This indicate that particulate matters concentration level has obvious seasonal patterns.

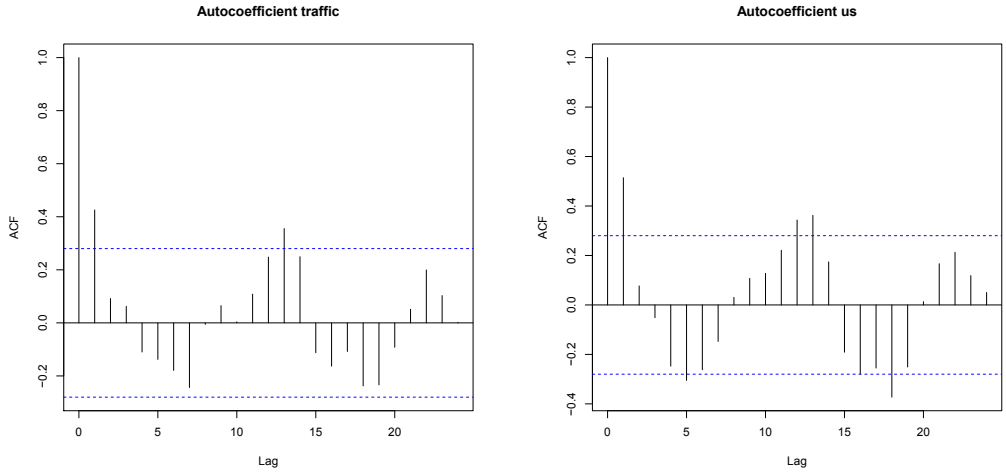

Figure 4.1.2: left: Autocorrelation Coefficients in traffic intensive area; right: Autocorrelation Coefficients in U.S. Embassy, see related quantlets: exploratory_analysis

And among all variables, $PM_{2.5}$ data collected from U.S. Embassy has the strongest autocorrelation with five out of 24 autocorrelation coefficients have reached significant interval. In comparison, particulate matter concentration data in traffic intensive area and urban area have least significant autocorrelation, which both only have two autocorrelation coefficients exceed significant line and they are at the same position(1st unit and twelfth unit).

Extreme values in data sets are maybe caused by extreme event. Hence, these outliers need to be treated and analyzed carefully. In the following table, dates with the top 10 highest measurements during whole time period are listed. February 15th, 2014 is marked as pink, on which day is the traditional Lantern Festival in China. The Lantern Festival falls on the 15th day of the 1st lunar month and this day is also the last day that fireworks are permitted(Beginning of fireworks permission is the day before Chinese New Year[13]). In Beijing, there will also be multiple firework shows in the city. Fireworks increase particulate matter concentration level extremely and the influence will expand for the following days[14]. Extreme values on January 1st, 2017 is marked as blue. This maybe share the same reason as February 15th, 2014. Because there are massive firework shows in the Beijing Olympic Park to celebrate New Year's Day[15].

[13]Since the New Year of 2005, selling and setting fireworks and firecrackers are only permitted in the following time period in Beijing City: 30/01/2014-14/02/2014; 18/02/2015-05/03/2015; 07/02/2016-22/02/2016; 27/01/2017-11/02/2017

[14]China Youth Daily `http://zqb.cyol.com/html/2014-02/16/nw.D110000zgqnb_20140216_5-02.htm`(translated by the author, retrieved on 05.08.2018)

[15]Global Times: `http://world.huanqiu.com/weinxingonghao/2017-01/9982948.html`(translated by the author, retrieved on 05.08.2018)

Table 3: Dates with top 10 highest values for each area

	urban	rural	traffic	U.S. Embassy
1	2015-12-25	2015-12-01	2015-12-01	2015-12-25
2	2015-12-01	2014-02-15	2015-12-25	2017-01-01
3	2017-01-01	2014-02-26	2014-02-15	2014-02-25
4	2017-01-04	2014-02-25	2014-10-09	2015-12-01
5	2014-02-15	2017-01-01	2014-02-25	2016-12-21
6	2016-12-21	2015-12-25	2017-01-01	2014-10-09
7	2016-12-20	2015-01-15	2016-12-21	2014-01-16
8	2014-01-16	2016-12-21	2014-10-10	2014-10-25
9	2014-02-25	2014-01-16	2014-10-25	2015-01-15
10	2015-11-30	2016-03-04	2017-01-04	2016-12-20

Table 3: Dates with top 10 highest values for each area, see related quantlets: 
exploratory_analysis

4.2 Factor Analysis

There are various possible reasons for particulate matters: traffic density, industrial production, public and household heating system using coal as power supply, meteorological variables(wind speed, wind direction, temperature et cetera). In this chapter an exploratory factor analysis will be carried out to define latent variables. *KMO* coefficients of all variables in both stations are greater than 0.5, which implies that data from different are suitable for factor analysis. And according to scree plot, we applied one-factor and two-factor analysis to grouped daily average value of particulate matters concentration.

Table 4: Factor Analysis

area	one-factor	two-factor	
	MR1	MR1	MR2
urban	0.994	0.995	0
rural	0.991	0.991	0
traffic	0.990	0.995	0
U.S. Embassy	0.974	0.973	0
Proportion Variance	0.975	0.977	0.003

Table 4: Factor Analysis, see related quantlets:  exploratory_analysis

Table 4 shows the result of factor analysis. The result shows that one-factor model explains to great extent of all areas. One-factor model has explained 97.5% of total variance in the data of all areas. Especially in the data of urban area, rural area and traffic intensive area, the one-factor has explained more that 99%. The first factor in two-factor model has great similarity with one-factor model. The first factor in the two-factor model has explained 97.7% of all data sets. The

second factor in the two-factor model has no statistically significant influence on any data sets.

This factor analysis shows that there are no significant difference factors influencing particulate matters concentration data across different areas in Beijing, which is surprising that traffic density and difference between urban and rural areas do not play a significantly different role in different areas. This could be caused by a overall high level of traffic density in whole Beijing area, so that the definition of "traffic intensive" is not effective. The latter could be explained that Beijing is a mega city and the distinguish between urban area and rural area is not obvious anymore. The phenomenon above could also result from one or multiple factors that influence all measurement area to a extreme big extent, that the effect of different traffic density and urbanization is concealed.

5 Discontinuity Test

In this chapter, we investigate whether the data published by BJMEMC(Beijing Municipal Environmental Monitoring Center) has anomalies around the critical values calculated from API Index, when $PM_{2.5}$ is the primary cause for this air pollution Index. To answer this question, discontinuity test proposed in McCrary (2008) in the context of regression discontinuity design is applied to daily average data of all 28 observation stations.

When $PM_{2.5}$ is the primary cause of the air pollution for the day, then the concentration level of the IAPI(Individual Air Pollution Index) of this pollutant will be considered as the API for this day. In the new ambient air quality standards *GB3095-2012*, API exceed 100 means that the air pollution level can cause damage to human health. This is also the critical value for "Blue Sky Day". So if there is a incentive for lowering $PM_{2.5}$ concentration levels until it is below this critical value. But when the pollutants concentration is too high, then the visibility of the day will be decreased and a potential data manipulation will has high risk to be discovered. Therefore, a intuitive data manipulation will occur in a interval over critical value but not too far. If such behavior occurs significantly often enough, then the distribution of the pollutant concentration will display a discontinuity around the cut-off. If there is no sign of data manipulation, the distribution of air pollutant concentrations should be continuous.

Brannlund, Runar and Lofgren(1996) suggested that emission of pollutants is subject to stochastic fluctuations and neither polluters or regulators have total control of that. Ghanem and Zhang(2014) has augmented this assumption that city control air pollution emission through regulating numerous polluters, even if each polluter's exogenous contribution to air quality is discrete, the aggregated air quality should be continuous. Therefore, the distribution of particulate matters concentration should satisfy the continuity assumption for the McCrary (2008)

test. It is also the reason to use $PM_{2.5}$ concentration data directly instead of API. Because API is a linear transformation of the highest pollutants' IAQI out of six different pollutants(Six pollutants are: SO_2, NO_2, PM_{10}, $PM_{2.5}$, CO, O_3)[16].

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + IAQI_{Lo}{}^{17} \tag{2}$$

In this form, BP_{Hi} and BP_{Lo} are the upper and lower boundaries of concentrations for each air quality level, and $IAQI_{Hi}$ and $IAQI_{Lo}$ are the corresponding upper and lower index classes.

The pollutant with the highest IAQI across six pollutants will be regarded as the daily primary pollutant and the corresponding IAQI will be published as the daily API.

$$API = max\{IAQI_1, IAQI_2, ...IAQI_n\} \tag{3}$$

So that API is a non-linear transformation of all six pollutants concentration and the distribution is non-continuous. This doesn't fulfill the requirement of McCrary-Discontinuity-Test and the credibility of the test resulted will be biased. A potential data manipulation could happen in the following positions: in the process of calculating daily average pollutant concentrations, at station level or city level. It could be caused by the means of data falsification, which is against the law or loopholes. Daily average value could also be lowered by simply throwing out extreme high values and subsume this date loss into equipment fault(in comparison with Ghanem, Zhang, 2014). Since 2013, the New Ambient Air Quality Standards

[16]Ambient Air Quality Standards *GB3095-2012*:`http://210.72.1.216:8080/gzaqi/Document/gjzlbz.pdf`(retrieved on 05.08.2018)

[17]Technical Regulation on Ambient Air Quality Index *HJ633-2012*(on trial) `https://web.archive.org/web/20130430001557/http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/jcgfffbz/201203/W020120410332725219541.pdf`(retrieved on 05.08.2018)

GB3095-2012 has increased the number of measurements from 12 to 20 for an effective monitoring for daily average value.

A potential data manipulation are more likely to occur near the cut-offs, because fine particulate matters are the main cause of bad visibility. Hence, actions at very high level of particulate matters will catch attention from both citizens and central government officials. Ghanem and Zhang(2014) stated that, "manipulation right around the cut-off is less likely to be detected because the difference in visibility and other weather conditions associated with air quality may be indiscernible between API values at 100- and 100+". So that it is less risky to manipulate data a little bit beyond the critical value, since it is less detectable without professional measure equipment. The following table shows API categories with corresponding $PM_{2.5}$ concentration values according to *GB3095-2012*. In the next step, cut-off at $75\mu g/m^3$ will be firstly applied to $PM_{2.5}$ concentration data for all stations. Because AQI at 100 is the critical value for a "Blue Sky Day"(when air pollution level is in the categories of "Excellent" or "Good", this day is considered as a "Blue Sky Day").

Table 5: Air Quality Index with corresponding Concentration Level

$IAQI_{PM2.5}$	$PM_{2.5}$ $(\mu g/m^3)$	Air Pollution Level	Air Pollution Category	Health Implications
0	0	1	Excellent	No health implications
50	35	2	Good	Some pollutants may slightly affect very few hypersensitive individuals
100	75	3	Lightly Polluted	Healthy people may experience slight irritations and sensitive individuals will be slightly affected to a larger extent
150	115	4	Moderately Polluted	Sensitive individuals will experience more serious conditions. The hearts and respiratory systems of healthy people may be affected
200	150	5	Heavily Polluted	Healthy people will commonly show symptoms. People with respiratory or hearts disease will be significantly affected and will experience reduced endurance in activities
300	250	6	Severely Polluted	Healthy people will experience reduced endurance in activities and may also show noticeably strong symptoms. Other illness may be triggered in healthy people. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid outdoor activities.
400	350			
500	500			

5.1 Test Design

The density test proposed by McCrary (2008) is to estimate of the log difference of the heights of the density of the target variable between the left and right limit at the cut-off: $\theta = \ln \lim_{r \downarrow c} f(r) - \ln \lim_{r \uparrow c} f(r) \equiv \ln f^+ - \ln f^-$

Intuition of this test is simple, that a potential manipulation at or multiple specific cut-offs will change the distribution of the data. And hence will cause a difference at the position of the cut-offs. From the size and positive or negative

sign of the test-result, a discontinuity in the running variable can be concluded.

First step of McCrary test, a histogram is produced. But in order to discretize the data, a bin size b need to be set, which reflects the estimated variance of the data: $\hat{b} = 2\hat{\sigma}n^{-1/2}$ where $\hat{\sigma}$ is the standard deviation of the running variable.

In the second step, the discretized data is used to estimate the left and right limit of the density function at the cut-off using the chosen bandwidth h.(in comparison with McCrary, 2008). In this theory, it is recommended that the ratio of bandwidth h and bin size b, a h/b shall be greater than 10. And McCrary's theory shows that any ratio a greater than 10 should not cause variance in the test result. For the sake of completeness. In this report, all bandwidths are calculated using selection calculation from McCrary (2008) automatically by using the *DCdensity* function in `rdd` package.

According to the p-value provided by the test result, it can be concluded whether there is evidence consistent with manipulation. The Null- Hypothesis of this test design is : there is no sorting in the running variable. When a p-value is smaller that the significant , therefore the Null- Hypothesis can be rejected. Because a effective manipulation should change the pollution level downwards, which would lead to a discontinuity at the cut-off, where the left limit is higher than the right limit. And the estimated θ will be accordingly negative. To be noticed that, θ need to be interpreted properly. Because it is the log difference of the density from the right side of the cut-offs to the left side of the cut-offs. It shows a difference in percentage, instead of directly indicating to what extent the manipulation has been done. So the scale of manipulation also depending on the position of the cut-offs in the data distribution.

5.2 Baseline Result

After the application of the McCrary test to daily average particulate concentration data for all 28 measurement stations in Great Beijing Area, the test result is illustrated in the following heatmap. Because of the low amount of data for the year 2013, so data from year 2013 is combined with data in year 2014 and test together. The Test result comprise two values: θ and P-value. θ indicates the log difference between the right limit of the density and the left limit of the density. When θ is negative, this means that the density on the left side of the cut-offs is higher than the right side of the cut-offs, which is in the point of our interest. P-value shows different levels of significance of the evidence for potential manipulation in the $PM_{2.5}$ data across all stations. P-values according to their distance the significance critical value are marked from dark red over pink to white.

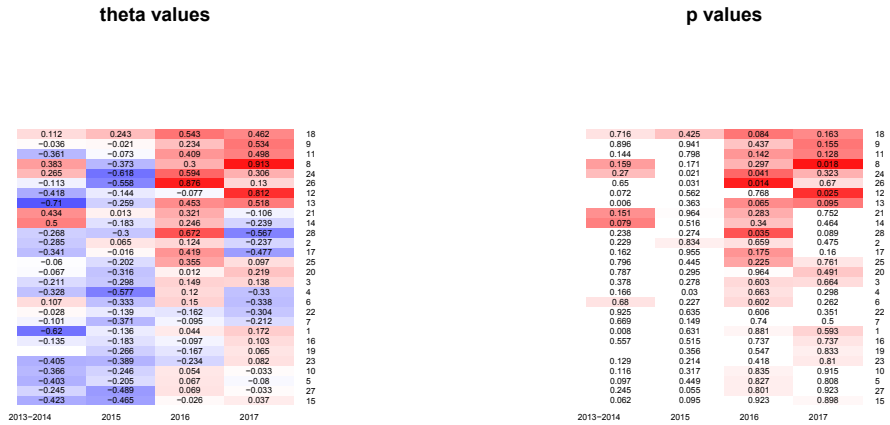

Figure 5.2.1: left: theta value of density test: "blue" for negative, "red" for positive; right: significance of test result, "red" for p-value in significant interval, see related quantlets:

exploratory_analysis

The McCrary test result shows 10 potential manipulation behaviors out of 112 tests. This test result suggests that 8.9% of our samples have discontinuity in $PM_{2.5}$ pollution data in the time period from 2013-2017. Table 6 exhibits the stations that are have statistically significant discontinuity around the cut-offs. More specifically, in five of all cases, the left limit of the density is significantly higher than the right limit, and the other half of all cases, the left limit of the density is significantly higher than the right limit. To be noticed that, in the perspective of local measurement stations, the latter behaviour is right in the opposite of their interest.

Table 6: Stations with Discontinuity in data

Station No.	Number of Discontinuity	Year of Discontinuity
1	1	2013-2014
4	1	2015
8	1	2017
12	1	2017
13	1	2013-2014
24	2	2015, 2016
26	2	2015, 2016
28	1	2016

Table 6: Stations with Discontinuity in data, see related quantlets:  exploratory_analysis

Table 7 shows that when data from single measurement stations are averaged as different areas and whole city, the test result for McCrary test shows no signs of discontinuity.

Table 7: Results for McCrary Test ($PM_{2.5}$)

	θ	P-value
urban	-0.1694	0.2681
rural	-0.0199	0.9052
traffic	0.1267	0.4216
whole city	-0.2092	0.1965

Table 7: Results for McCrary Test, see related quantlets:  exploratory_analysis

So according to the small scale of discontinuities and contradictory behaviour in the discontinuity test, an obvious evidence for a massive data manipulation is not shown in the test result. Ghanem and Zhang(2014) has applied the same test to data of PM_10 concentrations in 113 Chinese cities in the time period of 2001 to 2010 and suggest that 61 cities, 55% of all cities, reported dubious PM_10 pollution data and the manipulation level of Beijing has ranked number four among all cities. This difference shows that the new regulations rules of data measuring and reporting and stricter supervising and controlling methods for local government and related organizations has obviously reduced the data manipulation. But this conclusion only base on the discontinuity test and is under the continuity assumption that the hourly measurement data is trustworthy and distribute without any outsider effect.

5.3 Robustness Check and Caveats

As robustness check, the same density discontinuity test method is also applied to the $PM_{2.5}$ data collected and published by the U.S. Embassy. Because U.S. Embassy doesn't have the incentive to under-reporting particulate matters concentration level, so that under this assumption, there should be no discontinuity. And the test result(θ: -0.1408; P-value: 0.3835) of whole period(Dec. 2013 - Jun. 2017) shows that air pollution data from U.S. Embassy has no sign of discontinuity in McCrary test.

One caveat of the test result is that, the assumption of continuity of data in each year may be biased by the data size(about 360 data points). Because in the test of whole period(1491 data points) has no sign of discontinuity. Another caveat to this approach is that only the types of manipulation that lead to a discontinuity can be found out. For example, if all measurement equipment are set to reduce a certain amount of value from the real concentration level, this would not lead to a discontinuity because the data distribution remains the same, only with a lower mean value.

6 Conclusion

In this paper, we propose to apply descriptive data analysis and exploratory data analysis to the particulate matters($PM_{2.5}$)data of great Beijing area, to find out patterns in the time series with a total time period from Dec. 2013(which is the start of officially measuring and publishing of $PM_{2.5}$) to Jun. 2017.

Descriptive data analysis shows that during the whole time period yearly average level of $PM_{2.5}$ in all subgroups have been way over critical value of 35 $\mu g/m^3$. But there is trend of decrease shown in the air pollution level. And over half of all the daily average value of $PM_{2.5}$ are below critical value of 35 $\mu g/m^3$. Box plots have compared the air pollution situation among different areas, while similarities in the box-plots suggests that the distributions of air pollution data share many consistant characters.

Exploratory analysis shows very clear seasonal patterns are obvious in all areas. Winter and fall have overall higher air pollution than spring and summer. Extreme high values are highly correlated with some specific dates, which proves that these extreme values are often event-driven. Factor analysis has defined that, there is one main factor that contribute to the most of total variance. But in order to find out what could be the main factor, more variables like traffic density and meteorological variables should be included in this study.

In the part of discontinuity test, potential manipulation in the data is of our main interest. But the test result suggests very few evidence of discontinuity in the data sets. But we can not rule out the possible manipulation actions that will not lead to discontinuity.

7 Appendix

Table 8: Discontinuity Test Result for all 28 stations

	2013-2014		2015		2016		2017	
station No.	theta	P-value	theta	P-value	theta	P-value	theta	P-value
1	-0.62	0.008	-0.136	0.631	0.044	0.881	0.172	0.593
2	-0.285	0.229	0.065	0.834	0.124	0.659	-0.237	0.475
3	-0.211	0.378	-0.298	0.278	0.149	0.603	0.138	0.664
4	-0.328	0.166	-0.577	0.03	0.12	0.663	-0.33	0.298
5	-0.403	0.097	-0.205	0.449	0.067	0.827	-0.08	0.808
6	0.107	0.68	-0.333	0.227	0.15	0.602	-0.338	0.262
7	-0.101	0.669	-0.371	0.149	-0.095	0.74	-0.212	0.5
8	0.383	0.159	-0.373	0.171	0.3	0.297	0.913	0.018
9	-0.036	0.896	-0.021	0.941	0.234	0.437	0.534	0.155
10	-0.366	0.116	-0.246	0.317	0.054	0.835	-0.033	0.915
11	-0.361	0.144	-0.073	0.798	0.409	0.142	0.498	0.128
12	-0.418	0.072	-0.144	0.562	-0.077	0.768	0.812	0.025
13	-0.71	0.006	-0.259	0.363	0.453	0.065	0.518	0.095
14	0.5	0.079	-0.183	0.516	0.246	0.34	-0.239	0.464
15	-0.423	0.062	-0.465	0.095	-0.026	0.923	0.037	0.898
16	-0.135	0.557	-0.183	0.515	-0.097	0.737	0.103	0.737
17	-0.341	0.162	-0.016	0.955	0.419	0.175	-0.477	0.16
18	0.112	0.716	0.243	0.425	0.543	0.084	0.462	0.163
19	NA	NA	-0.266	0.356	-0.167	0.547	0.065	0.833
20	-0.067	0.787	-0.316	0.295	0.012	0.964	0.219	0.491
21	0.434	0.151	0.013	0.964	0.321	0.283	-0.106	0.752
22	-0.028	0.925	-0.139	0.635	-0.162	0.606	-0.304	0.351
23	-0.405	0.129	-0.389	0.214	-0.234	0.418	0.082	0.81

Table 8: Discontinuity Test Result for all 28 stations

	2013-2014		2015		2016		2017	
24	0.265	0.27	-0.618	0.021	0.594	0.041	0.306	0.323
25	-0.06	0.796	-0.202	0.445	0.355	0.225	0.097	0.761
26	-0.113	0.65	-0.558	0.031	0.876	0.014	0.13	0.67
27	-0.245	0.245	-0.489	0.055	0.069	0.801	-0.033	0.923
28	-0.268	0.238	-0.3	0.274	0.672	0.035	-0.567	0.089

References

[1] Andrews, Steven. (2008a). Beijing Plays Air Quality Games. *Far Eastern Economic Review July/August 2008*:53-57.

[2] Andrews Steven.(2008b). Inconsistencies in air quality metrics: 'Blue Sky' days and PM10,concentrations in Beijing. *Environmental Research Letters 3 (3):034009.*

[3] Brannlund, Lofgren. (1996). Emission Standards and Stochastic Waste Load. *Land Economics 72 (2)*:218–230.

[4] Chen et al.(2012). Gaming in Air Pollution Data? Lessons from China.*The B.E. Journal of Economic Analysis Policy (Advances) 13 (3):Article 2.*

[5] Ghanem, Zhang.(2014). Effortless Perfection: Do Chinese cities manipulate air pollution data?. *Journal of Environmental Economics and Management, vol 68(2), ISSN0095-0696*: 203-225.

[6] Guo, Zhang. (2009). Correlation between PM concentrations and aerosol optical depth in eastern China. *Atmospheric Environment 43(37)*:5876-5886.

[7] Matus et al.(2012). Health damages from air pollution in China. *Global. Environmental Change 22 (1)*:55-66.

[8] McCrary, Justin. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142 (2)*:698–714.

[9] OECD. (2014) *The cost of air pollution: health impacts of road transport.* OECD Publishing, Paris.

[10] Pope, Dockery. (2006). Health effects of fine particulate air pollution: lines that connect. *Journal of the Air Waste Management Association 56(6)*:709–742.

[11] Schwartz et al. (1996). Is daily mortality associated specifically with fine particles? *Journal of the Air Waste Management Association 46*: 927–939.

[12] Shumway Robert H., Stoffer David S.(2011). *Time Series Analysis and Its Applications with R Examples*, Springer.

[13] WHO. (2006). *WHO air quality guidelines for particulate matter, ozone nitrogen dioxide and sulfur dioxide*. World Health Organisation (WHO), Geneva.

[14] Zheng, Kahn. (2013). Understanding China's urban pollution dynamics. *Journal of Economic Literature 51(3)*:731–772.