PM 2.5 in Beijing

Kang Fu, Dandan Feng, Xinning Chu, Ashley Hall

Background - As Mentioned Previously

PM2.5 are dangerous particles of pollutants that are less than 2.5 microns in diameter - including combustion particles, organic compounds, metals, etc.

RESPIRATORY
Insoluble part accumulates at the alveolus of the lungs, causing inflammation

BLOOD
Causes blood toxicity, blood coagulation abnormalities, and can trigger heart disease

CARDIOVASCULAR
Causes cardiotoxicity and also causes severe irritation to the autonomic nervous system, which regulates the activity of the heart muscle

REPRODUCTIVE
Placental blood toxicity which leads to direct harm to fetus, intrauterine growth, and low birth weight of babies

Beijing & Our Objectives

Beijing, specifically, was asking to bring PM2.5 concentration down to around 60 micrograms per cubic meter

1. Has Beijing's PM2.5 levels increased or decreased - air quality getting better or worse?
2. Has Beijing hit the goal of bringing PM2.5 concentrations down to around 60 micrograms per cubic meters?
3. What will Beijing's Air Quality look like in 10/15 years?
4. Suggestions on policies and procedures to further address PM2.5 concerns

Introduction to PM2.5 Estimation Method

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Data Source

- The U.S. Department of State Data: http://www.stateair.net/web/historical/1/1.html (2008 - June 2017)
- AQI study: https://www.aqistudy.cn/historydata/monthdata.php?city=北京 (July 2017 - March 2019)

Year | Month | Day | Hour | Value (PM2.5) | Unit (μg/m³) | QC Name (Valid or Missing)

Exploratory Data Analysis

1. Data pre-processing:

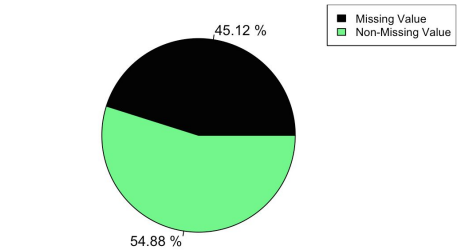
Old dataset (Data from the U.S. Embassy) & New dataset (Data from the AQI study)

For the old data set (Data from U.S. Embassy)

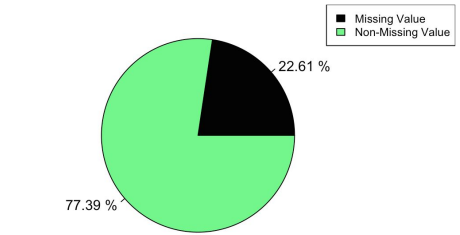
1. Check the missing values and delete them.
2. Calculate the daily max and daily average pm2.5 value using the original hourly data.
3. Calculate the monthly average pm2.5 value.

Hourly missing values in the old dataset

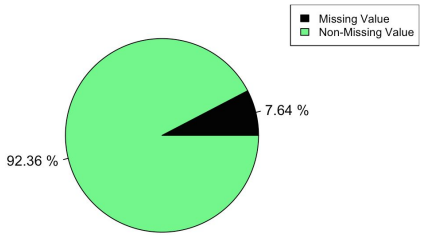
The percentage of Hourly Missing Value in 2008



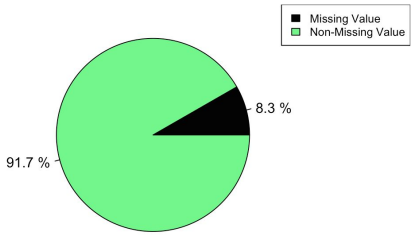
The percentage of Hourly Missing Value in 2009



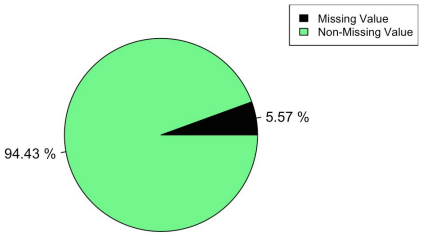
The percentage of Hourly Missing Value in 2010



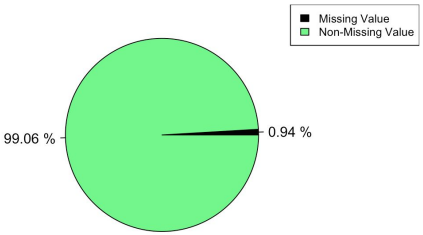
The percentage of Hourly Missing Value in 2011



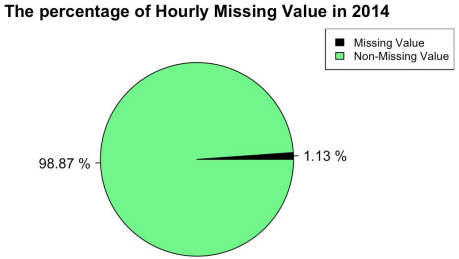
The percentage of Hourly Missing Value in 2012

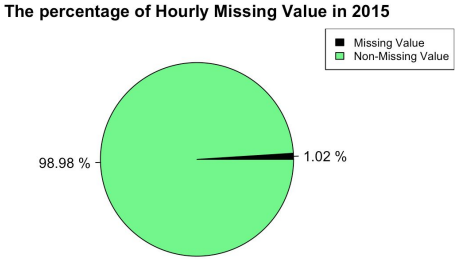


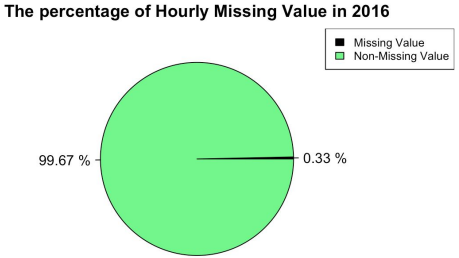
The percentage of Hourly Missing Value in 2013

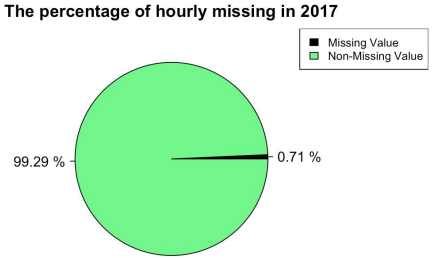


Hourly missing values in the old dataset





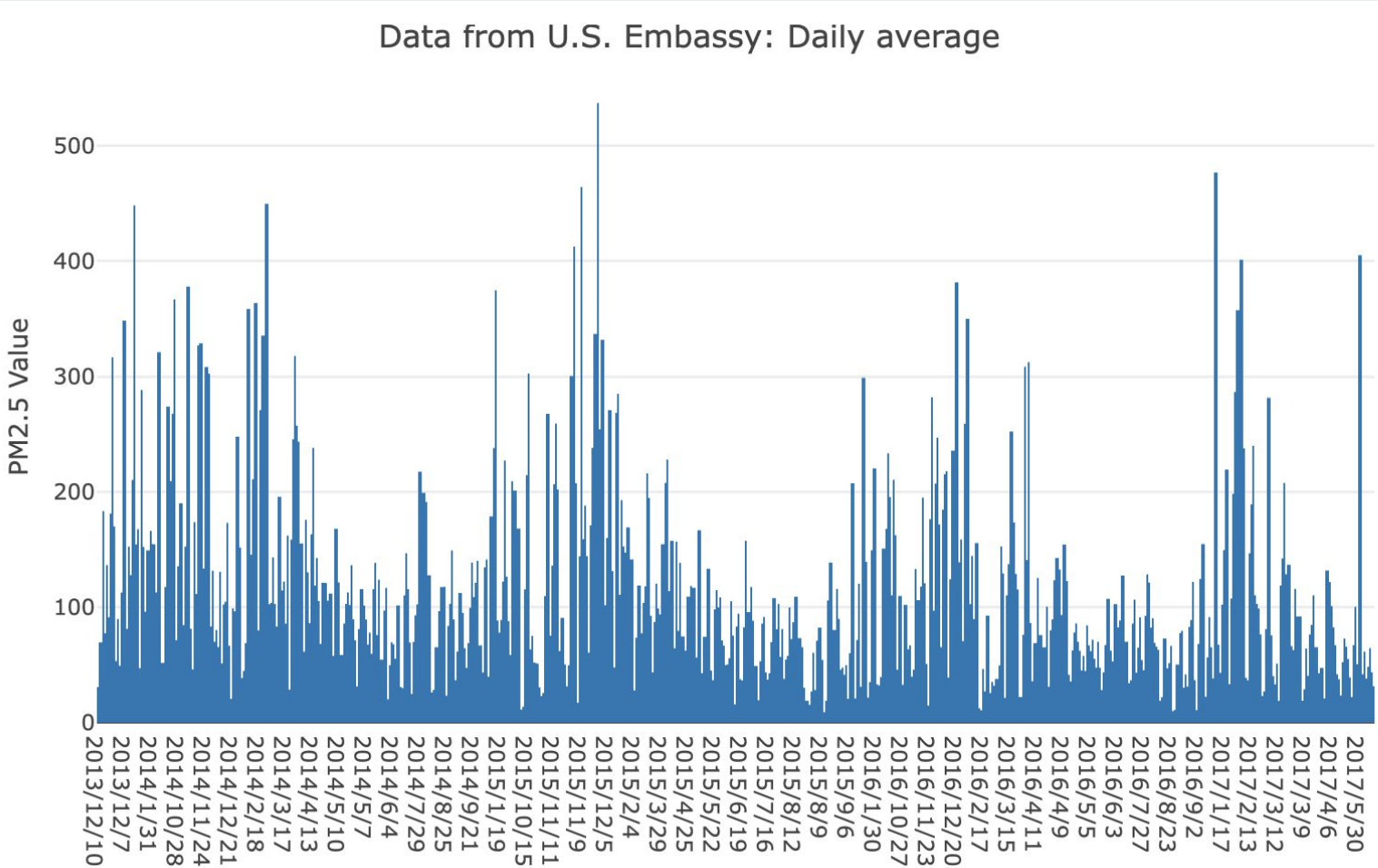


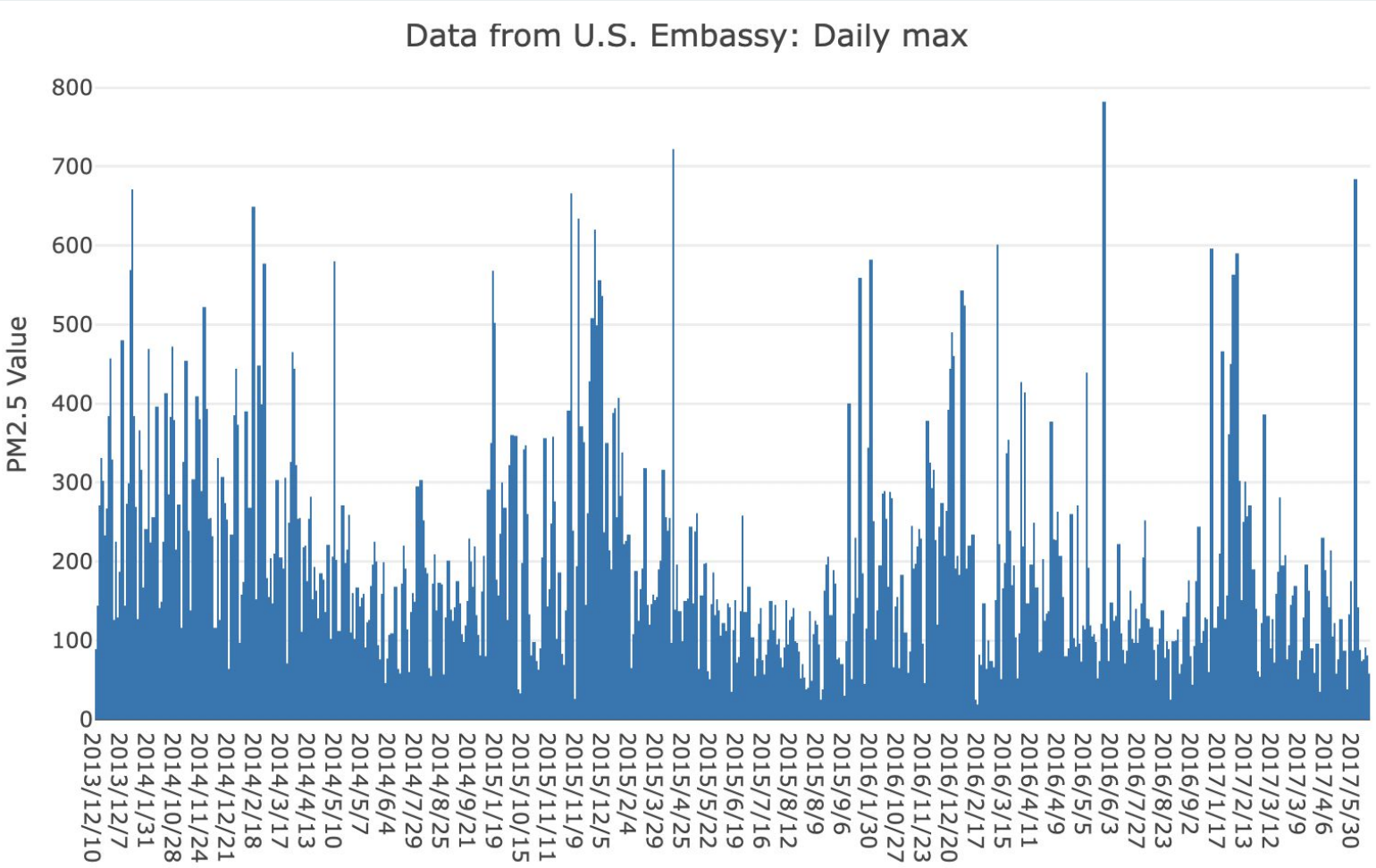


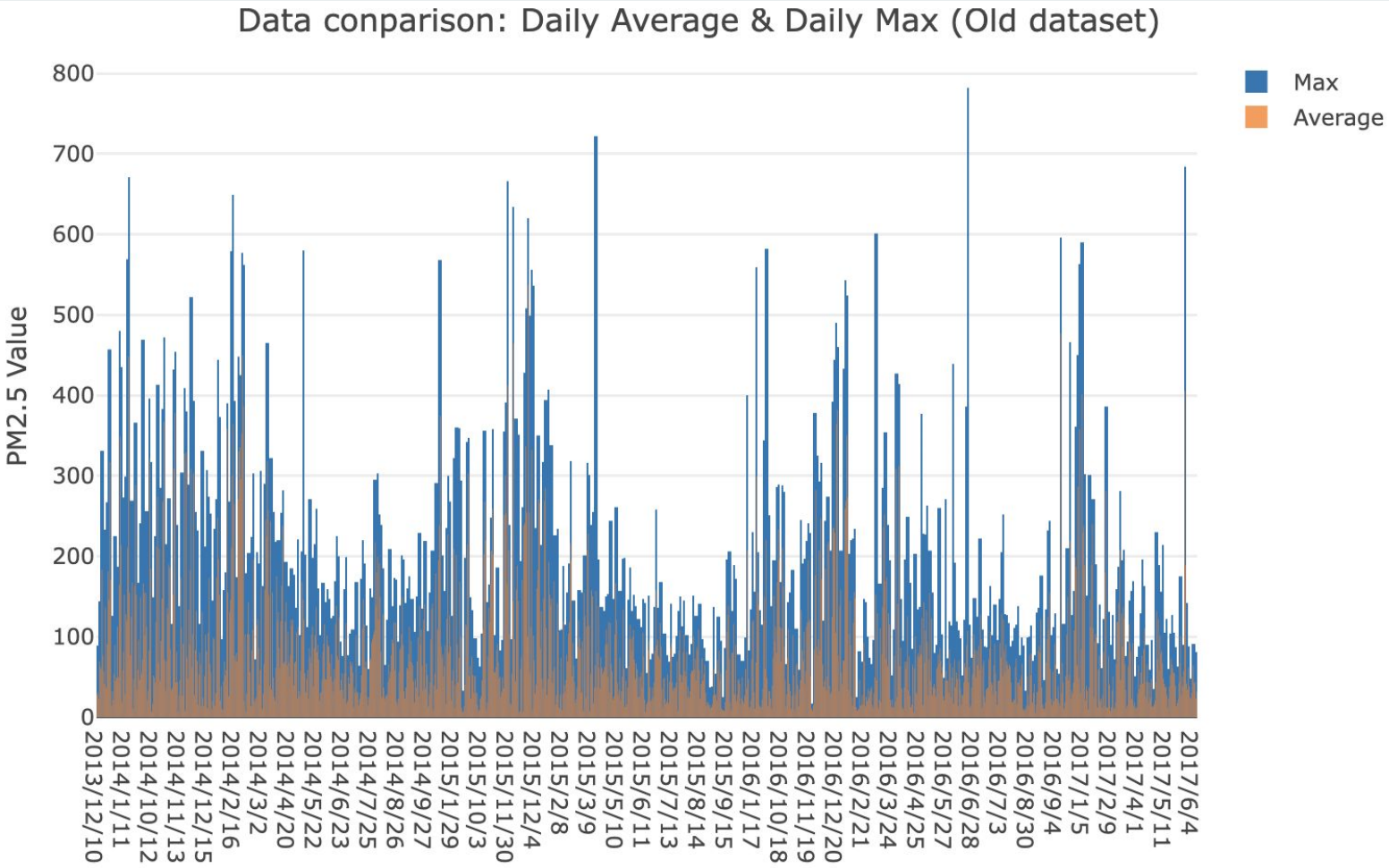
For the following reasons:

1. The ratios of missing values in 2008 and 2009 are too high.
2. 2008 & 2009 are far away from now.

We've decided to use the data from 2010/1/1 to 2017/6/9 (For the old dataset only contains the data from January 1st to June 9th in 2017) in the old data set.


Data from U.S. Embassy: Daily average
PM2.5 Value
500
400
300
200
100
0
2013/12/10
2013/12/7
2014/1/31
2014/10/28
2014/11/24
2014/12/21
2014/2/18
2014/3/17
2014/4/13
2014/5/10
2014/5/7
2014/6/4
2014/7/29
2014/8/25
2014/9/21
2015/1/19
2015/10/15
2015/11/11
2015/11/9
2015/12/5
2015/2/4
2015/3/29
2015/4/25
2015/5/22
2015/6/19
2015/7/16
2015/8/12
2015/8/9
2015/9/6
2016/1/30
2016/10/27
2016/11/23
2016/12/20
2016/2/17
2016/3/15
2016/4/11
2016/4/9
2016/5/5
2016/6/3
2016/7/27
2016/8/23
2016/9/2
2017/1/17
2017/2/13
2017/3/12
2017/3/9
2017/4/6
2017/5/30


Data from U.S. Embassy: Daily max
PM2.5 Value
800
700
600
500
400
300
200
100
0
2013/12/10
2013/12/7
2014/1/31
2014/10/28
2014/11/24
2014/12/21
2014/2/18
2014/3/17
2014/4/13
2014/5/10
2014/5/7
2014/6/4
2014/7/29
2014/8/25
2014/9/21
2015/1/19
2015/10/15
2015/11/11
2015/11/9
2015/12/5
2015/2/4
2015/3/29
2015/4/25
2015/5/22
2015/6/19
2015/7/16
2015/8/12
2015/8/9
2015/9/6
2016/1/30
2016/10/27
2016/11/23
2016/12/20
2016/2/17
2016/3/15
2016/4/11
2016/4/9
2016/5/5
2016/6/3
2016/7/27
2016/8/23
2016/9/2
2017/1/17
2017/2/13
2017/3/12
2017/3/9
2017/4/6
2017/5/30


Data conparison: Daily Average & Daily Max (Old dataset)
PM2.5 Value
800
700
600
500
400
300
200
100
0
Max
Average
2013/12/10
2014/1/11
2014/10/12
2014/11/13
2014/12/15
2014/2/16
2014/3/2
2014/4/20
2014/5/22
2014/6/23
2014/7/25
2014/8/26
2014/9/27
2015/1/29
2015/10/3
2015/11/30
2015/12/4
2015/2/8
2015/3/9
2015/5/10
2015/6/11
2015/7/13
2015/8/14
2015/9/15
2016/1/17
2016/10/18
2016/11/19
2016/12/20
2016/2/21
2016/3/24
2016/4/25
2016/5/27
2016/6/28
2016/7/3
2016/8/30
2016/9/4
2017/1/5
2017/2/9
2017/4/1
2017/5/11
2017/6/4

For the new data set (Data from AQI study)

Original data from AQI study:

Two kinds of data are provided: Daily & Monthly

(From 2013/12/02 to 2019/4/16)

Monthly data:

月份	AQI	范围	质量等级	PM2.5	PM10	SO2	CO	NO2	O3
2013-12	100	23~291	轻度污染	73	97	37	1.73	56	38
2014-01	125	26~402	轻度污染	94	123	51	1.948	65	37

Daily data:

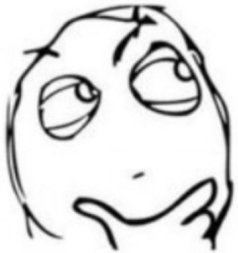
日期	AQI	质量等级	PM2.5	PM10	SO2	CO	NO2	O3_8h
2013-12-02	142	轻度污染	109	138	61	2.6	88	11
2013-12-03	86	良	64	86	38	1.6	54	45

Missing value in the AQI data set:

Only one missing value:

2016-04-15	80	良	59	107	16	1	58	91
2016-04-17	46	优	16	35	2	0.3	16	91

Did something big happened in Beijing on April 16, 2016?

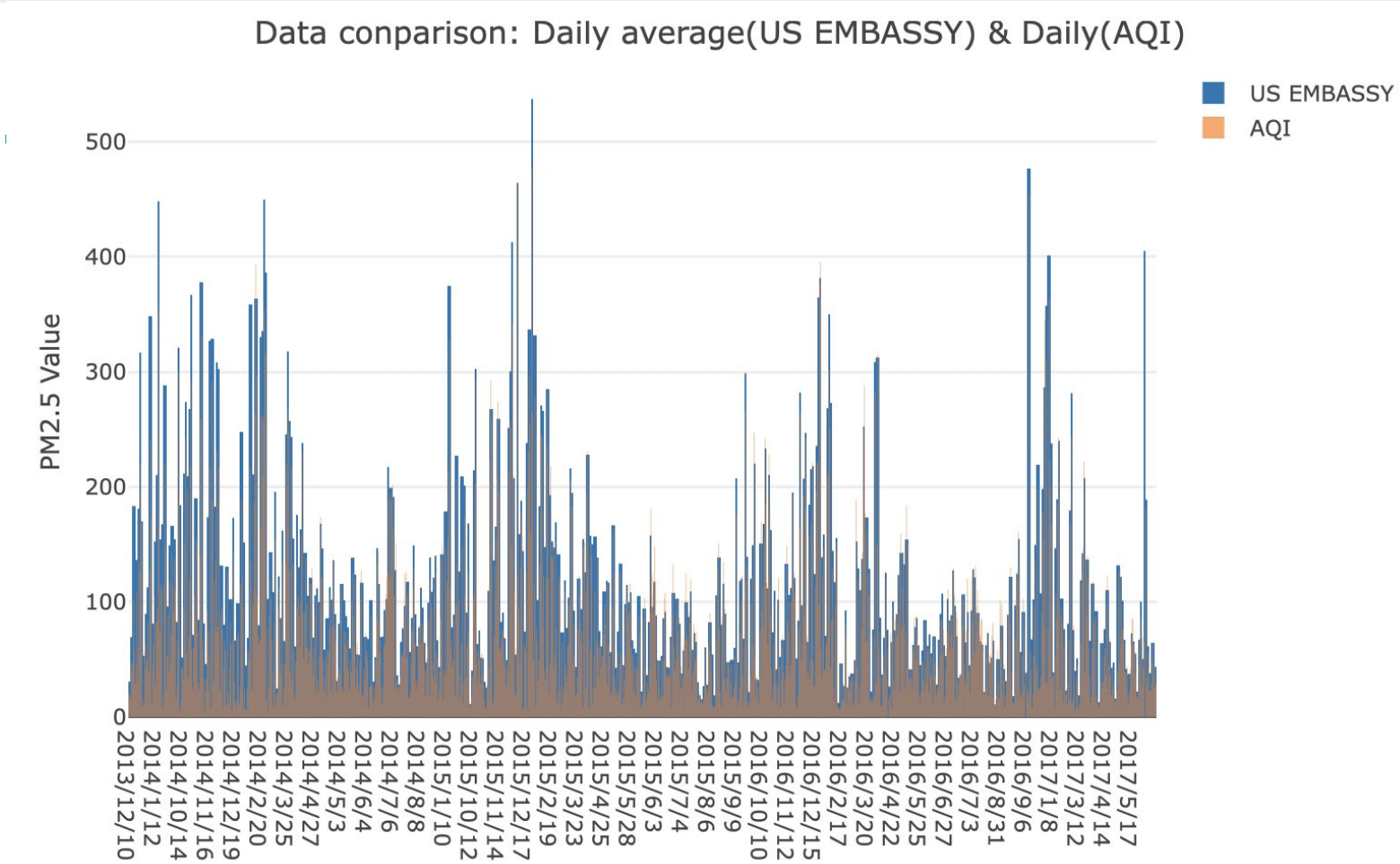


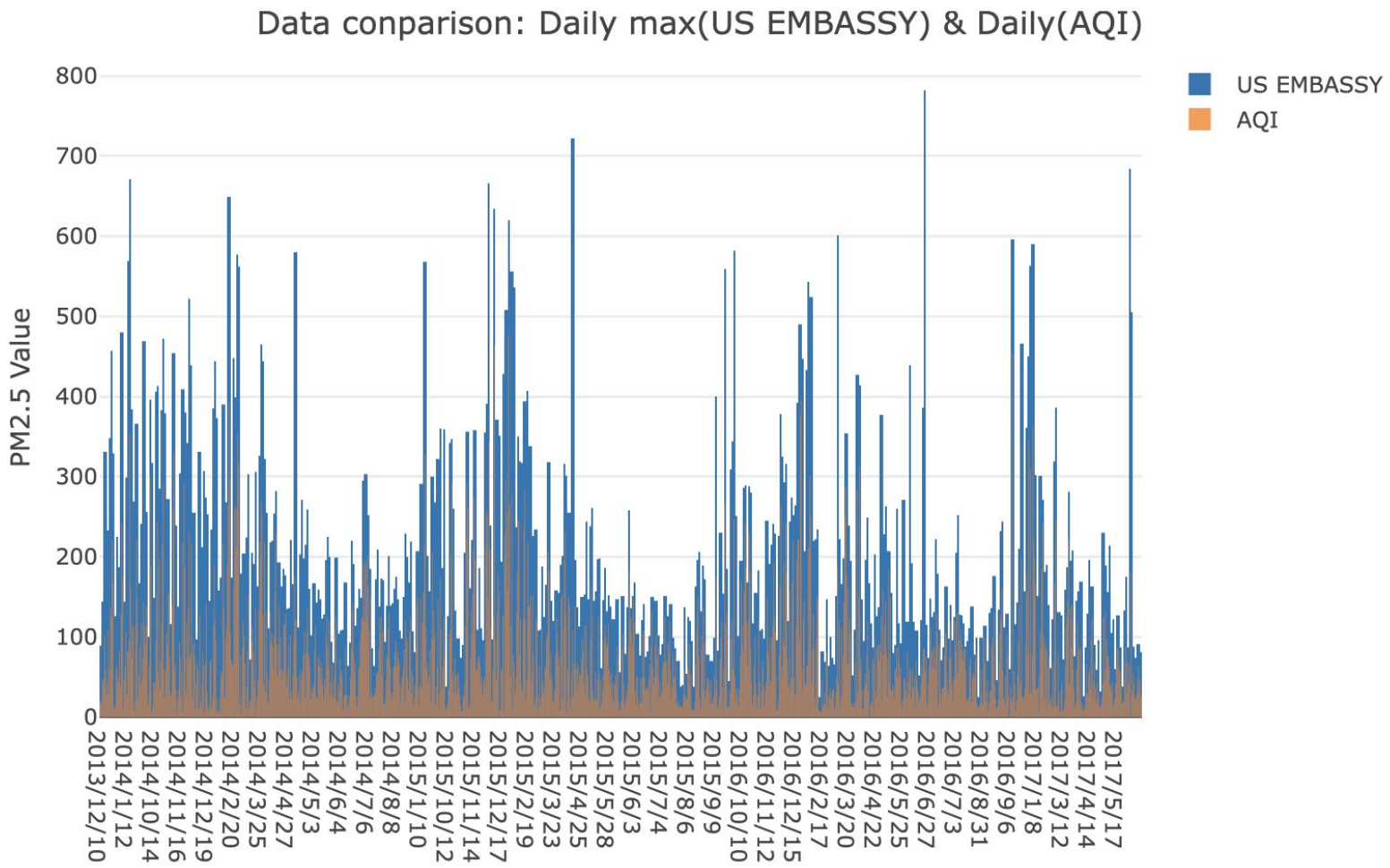
Exploratory Data Analysis

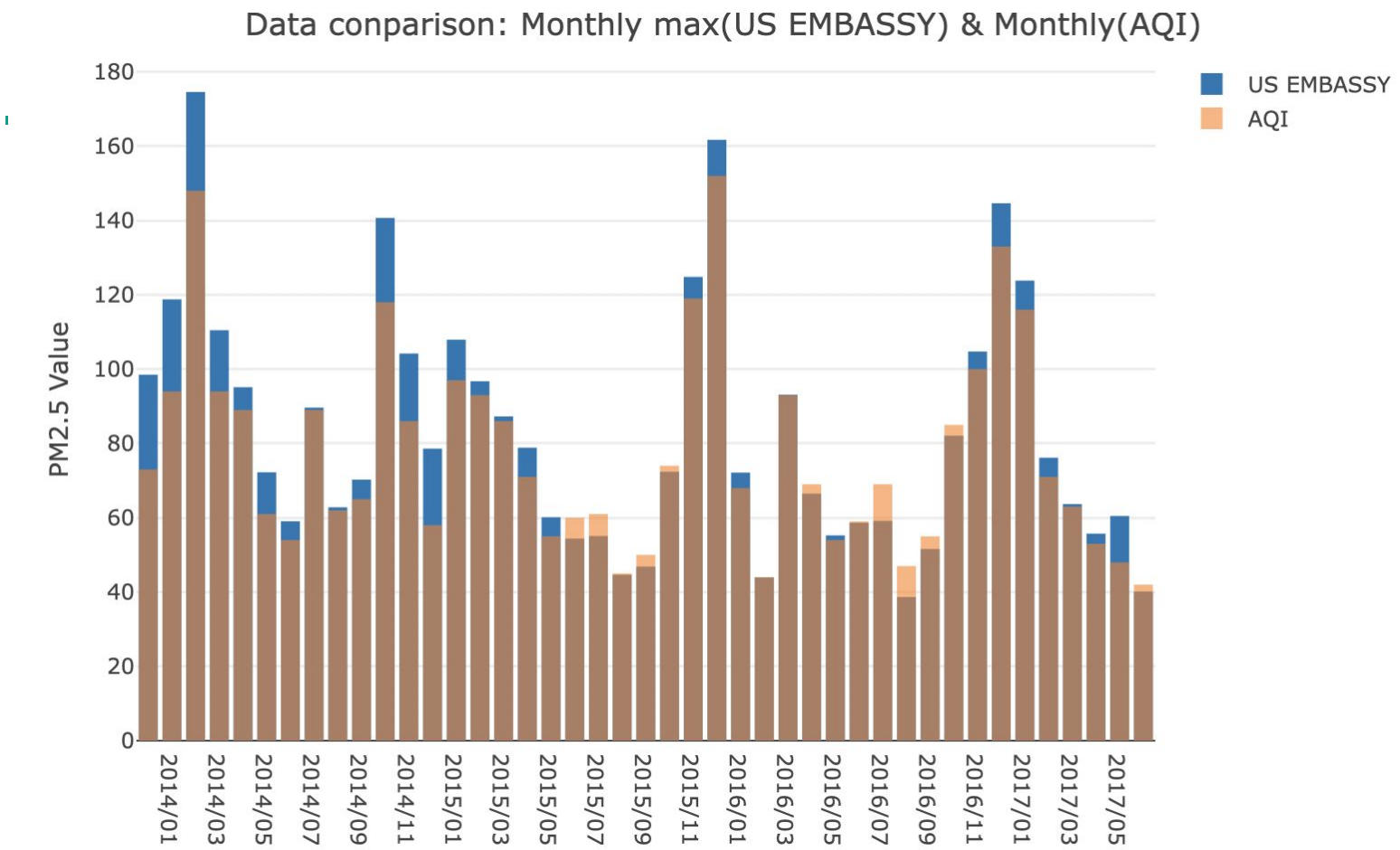
2. Relationship between the old and the new dataset:

Visualize & Compare statistically

Visualization:


Data conparison: Daily average(US EMBASSY) & Daily(AQI)
US EMBASSY
AQI
PM2.5 Value
500
400
300
200
100
0
2013/12/10
2014/1/12
2014/10/14
2014/11/16
2014/12/19
2014/2/20
2014/3/25
2014/4/27
2014/5/3
2014/6/4
2014/7/6
2014/8/8
2015/1/10
2015/10/12
2015/11/14
2015/12/17
2015/2/19
2015/3/23
2015/4/25
2015/5/28
2015/6/3
2015/7/4
2015/8/6
2015/9/9
2016/10/10
2016/11/12
2016/12/15
2016/2/17
2016/3/20
2016/4/22
2016/5/25
2016/6/27
2016/7/3
2016/8/31
2016/9/6
2017/1/8
2017/3/12
2017/4/14
2017/5/17


Data conparison: Daily max(US EMBASSY) & Daily(AQI)
US EMBASSY
AQI
PM2.5 Value
800
700
600
500
400
300
200
100
0
2013/12/10
2014/1/12
2014/10/14
2014/11/16
2014/12/19
2014/2/20
2014/3/25
2014/4/27
2014/5/3
2014/6/4
2014/7/6
2014/8/8
2015/1/10
2015/10/12
2015/11/14
2015/12/17
2015/2/19
2015/3/23
2015/4/25
2015/5/28
2015/6/3
2015/7/4
2015/8/6
2015/9/9
2016/10/10
2016/11/12
2016/12/15
2016/2/17
2016/3/20
2016/4/22
2016/5/25
2016/6/27
2016/7/3
2016/8/31
2016/9/6
2017/1/8
2017/3/12
2017/4/14
2017/5/17


Data conparison: Monthly max(US EMBASSY) & Monthly(AQI)
US EMBASSY
AQI
PM2.5 Value
180
160
140
120
100
80
60
40
20
0
2014/01
2014/03
2014/05
2014/07
2014/09
2014/11
2015/01
2015/03
2015/05
2015/07
2015/09
2015/11
2016/01
2016/03
2016/05
2016/07
2016/09
2016/11
2017/01
2017/03
2017/05

From the plots, we can see that the daily data in the AQI dataset is basically coincide with the daily average data in the U.S. Embassy dataset.

Statistical methods to strengthen our points:

1. Pearson correlation:

A **Pearson correlation** is a number between -1 and 1 that indicates the extent to which two variables are linearly related. The **Pearson correlation** is also known as the “product moment **correlation** coefficient” (PMCC) or simply “**correlation**”.

Pearson correlation value between the U.S. Embassy data and the AQI data according to each Year (2014, 2015, 2016, 2017) :

Year	Pearson Correlation
2014	0.9636894
2015	0.9841159
2016	0.9836109
2017	0.9034997

Highly correlated!

Statistical methods to strengthen our points:

2. Two sample t-test:

The two–sample *t*–test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application is to test if a new process or treatment is superior to a current process or treatment.

Year	P-Value for two sample t-test
2014	0.1792
2015	0.6578
2016	0.9911
2017	0.5026

No significant difference between the data from U.S. Embassy and AQI Study!

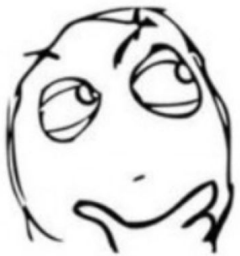
Thinkings behind the relationship of the old and new dataset:

1. For the former results showed that there is no significant difference between these two datasets, we can use the AQI data after 2017/6/9 to check the reliability of our former ARIMA prediction model (Which is built by using the old dataset that only contains the data from 2010/1/1 to 2017/6/9).
2. For the former results showed that there is no significant difference between these two datasets, we can link these two datasets together and get a new dataset that contains the data from 2010/1/1 to 2019/4/16. Then, use this new dataset to build a new ARIMA model.

Exploratory Data Analysis

3. Conduct some two sample t-tests to explore the possible variation trend of the PM2.5 value according to Years.

Year	P-Value	Estimated mean: (x,y)
2014(x) vs 2015(y)	0.009186(**)	(97.83859, 82.54337)
2015(x) vs 2016(y)	0.06692	(82.54337, 72.89106)
2016(x) vs 2017(y)	0.003919(**)	(72.89106, 57.37808)
2017(x) vs 2018(y)	0.02834(*)	(57.37808,49.44658)



Decreasing trend?

Next step:

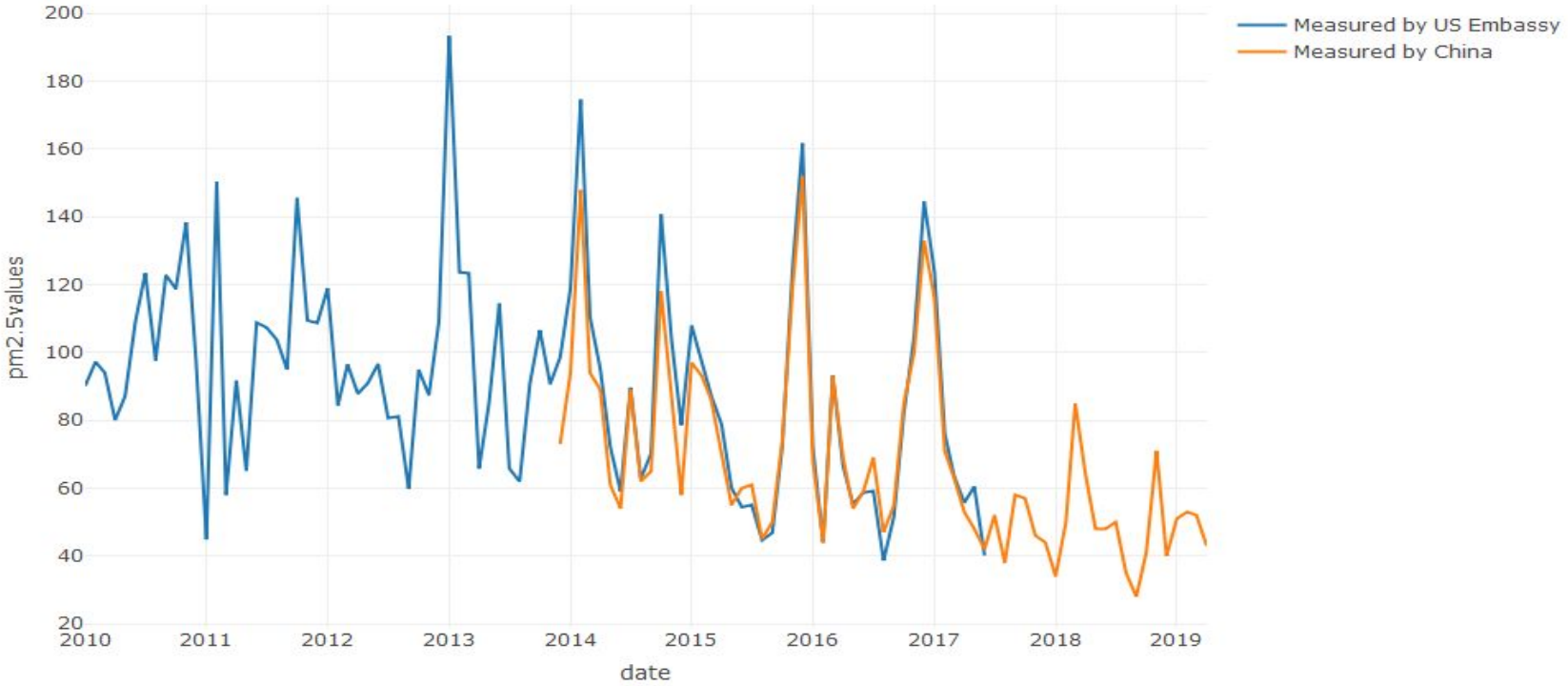
Use the ARIMA model to get more reliable results!

Overall PM2.5 Concentration Trend Analysis

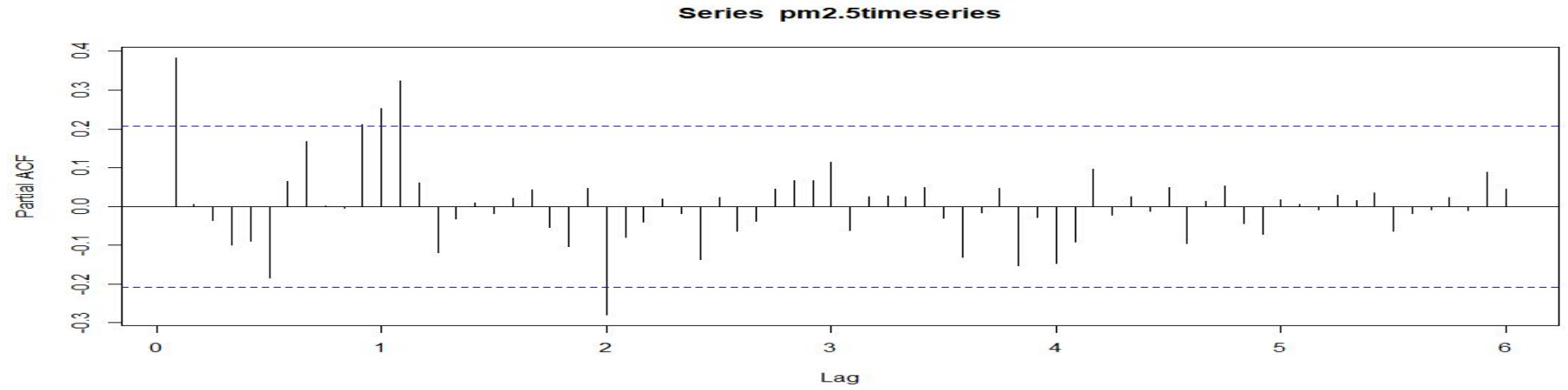
Methods:

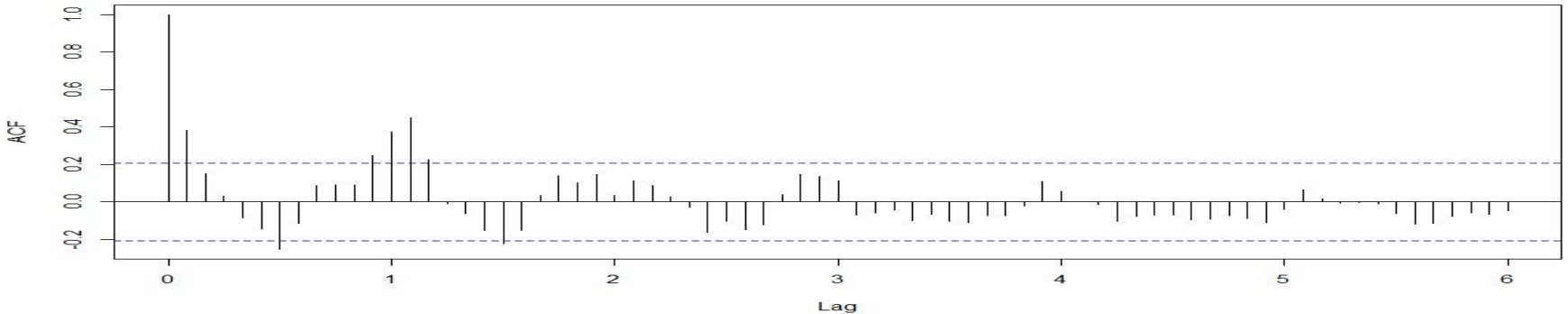
1. Making plots (time plot, ACF/PACF plots)
2. Stationary tests: Augmented Dickey-Fuller test; Phillips-Perron test
3. Build the ARIMA / SARIMA model
4. Diagnostics and forecasting

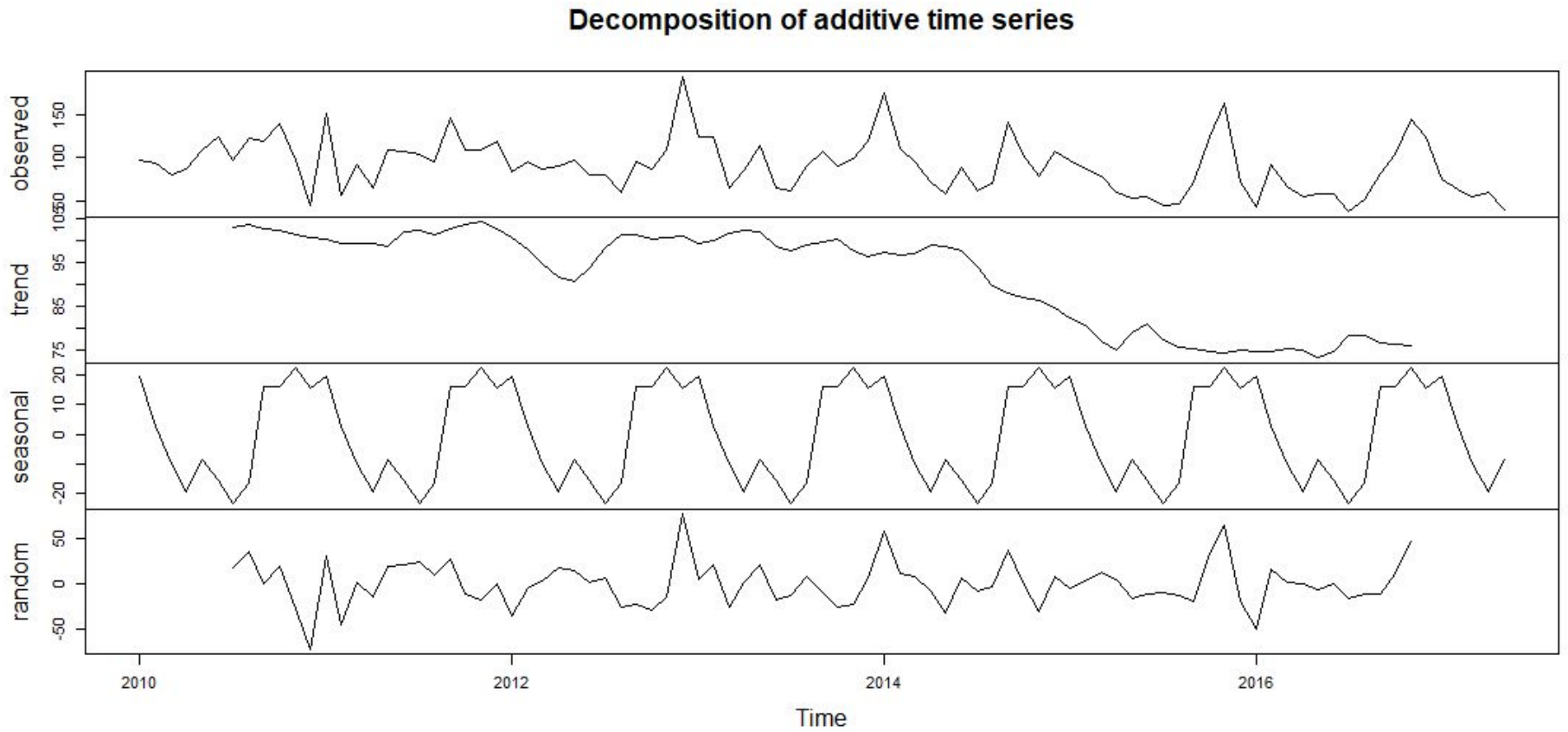
R packages used:
astsa / tseries / forecast / aTSA


Measured by US Embassy
Measured by China
pm2.5values
date
200
180
160
140
120
100
80
60
40
20
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

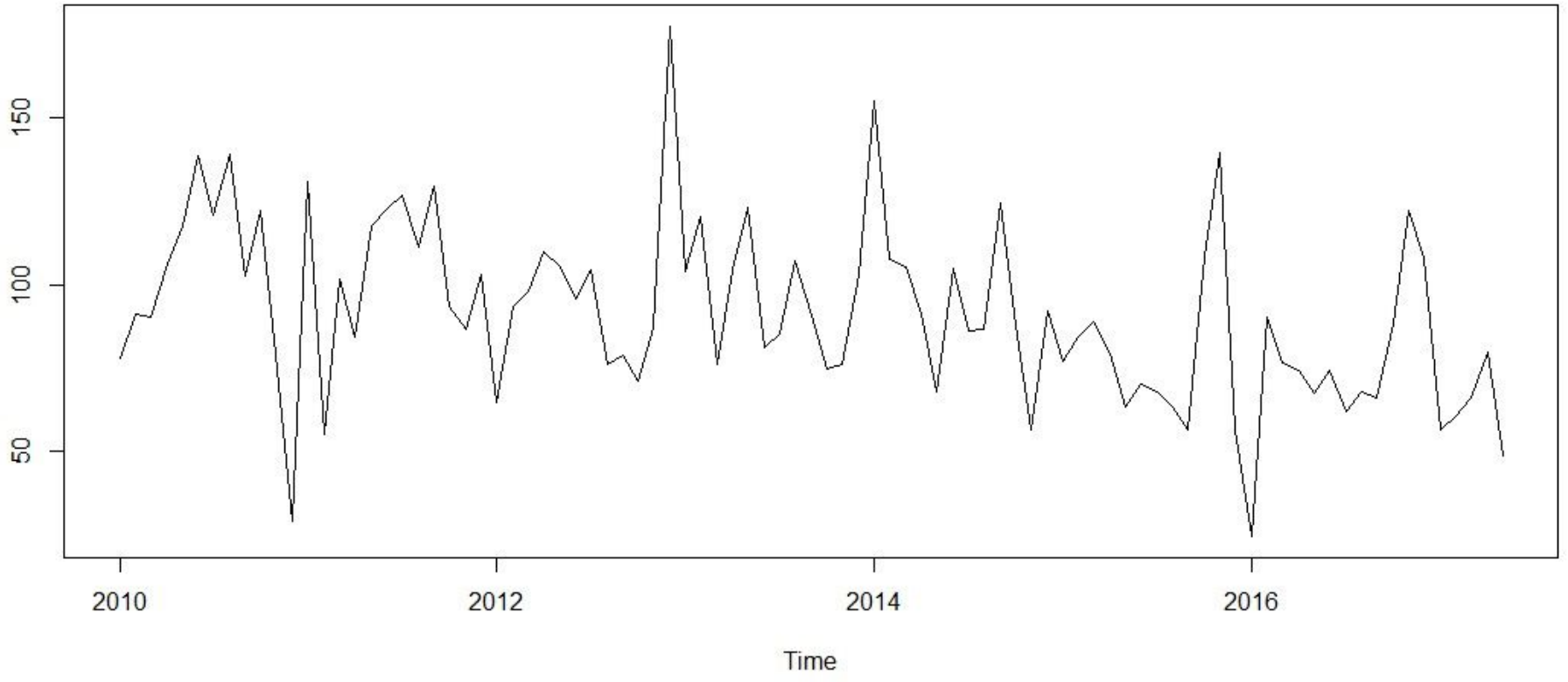
Time Series Analysis






Decomposition of additive time series
observed
trend
seasonal
random
2010
2012
2014
2016
Time

Time plot of pm2.5 values without seasonal factor



Results of Stationary Tests

```
> adf.test(pm2.5noseason1)

        Augmented Dickey-Fuller Test

data:  pm2.5noseason1
Dickey-Fuller = -5.0271, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(pm2.5noseason1) : p-value smaller than printed p-value
> pp.test(pm2.5noseason1)

        Phillips-Perron Unit Root Test

data:  pm2.5noseason1
Dickey-Fuller Z(alpha) = -71.254, Truncation lag parameter = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In pp.test(pm2.5noseason1) : p-value smaller than printed p-value
```

Parameters Estimation:

Model: (0,1,1)(0,0,1)[12]

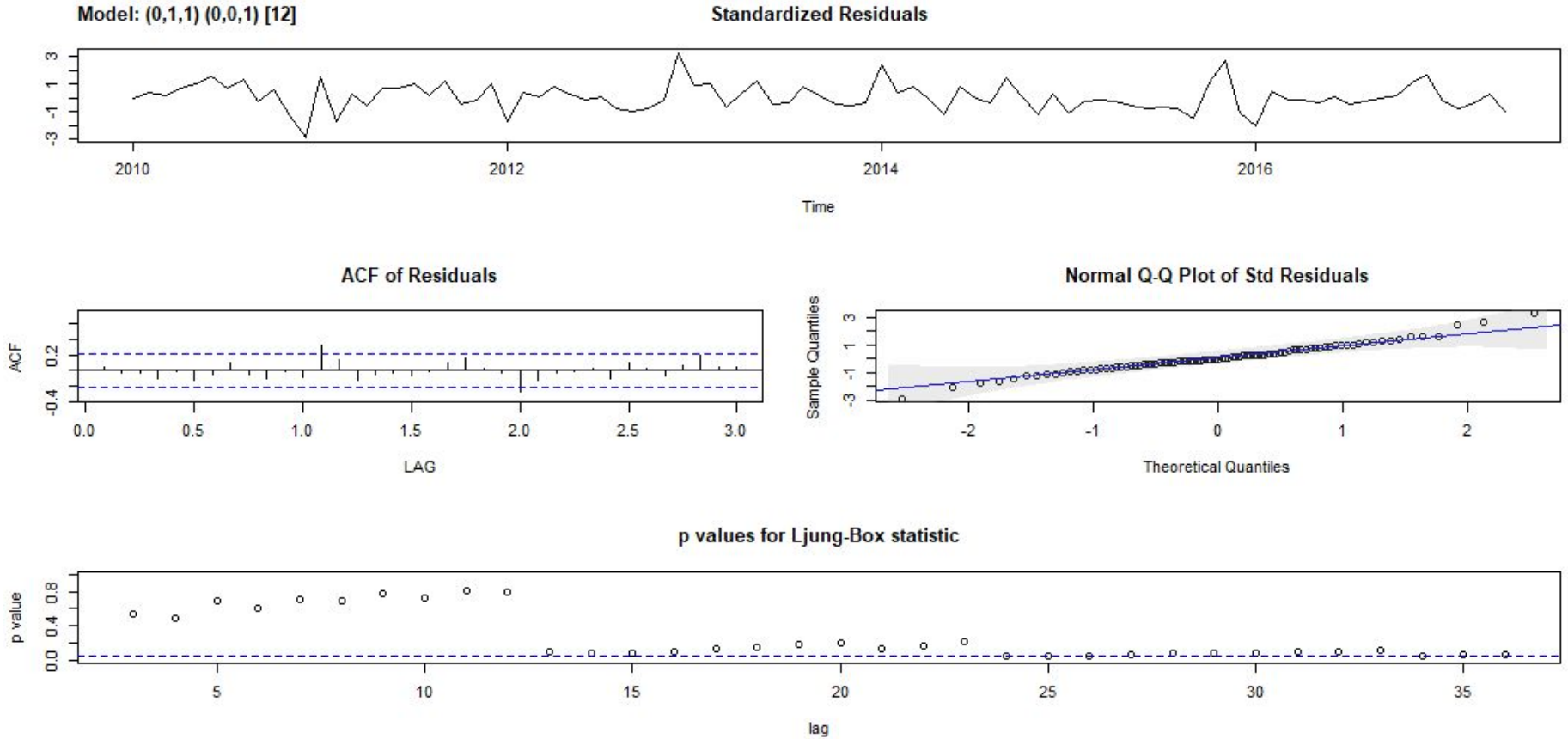
```
Coefficients:
          ma1     sma1  constant
      -1.0000  0.4949   -0.3593
s.e.   0.0309  0.1154    0.1435
```

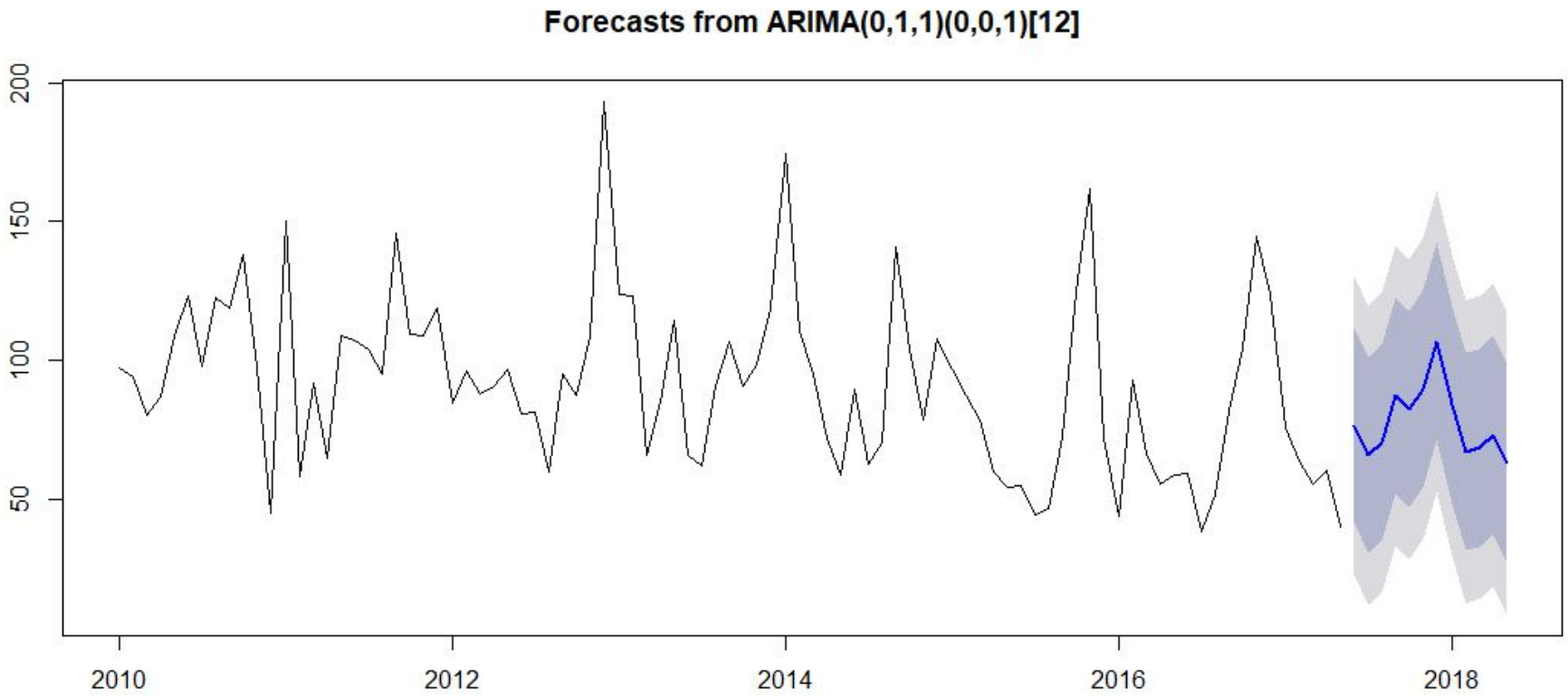
```
$ttable
         Estimate     SE  t.value p.value
ma1       -1.0000 0.0309 -32.3951  0.0000
sma1       0.4949 0.1154   4.2893  0.0000
constant  -0.3593 0.1435  -2.5031  0.0142

$AIC
[1] 7.612872

$AICc
[1] 7.640694

$BIC
[1] 6.696758
```


Model: (0,1,1) (0,0,1) [12]
Standardized Residuals
-3 -1 1 3
2010
2012
2014
2016
Time
ACF of Residuals
ACF
-0.4 0.2
0.0 0.5 1.0 1.5 2.0 2.5 3.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-3 -1 1 3
-2 -1 0 1 2
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0 0.4 0.8
5 10 15 20 25 30 35
lag


Forecasts from ARIMA(0,1,1)(0,0,1)[12]
200
150
100
50
2010
2012
2014
2016
2018

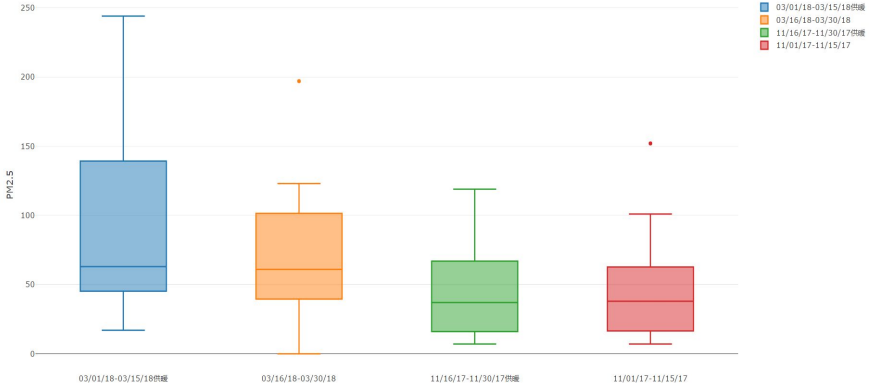
Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95	True Value
Jul 2017	76.65289	41.61452	111.69126	23.066327	130.2395	118
Aug 2017	65.71653	30.62840	100.80466	12.053872	119.3792	85
Sep 2017	70.36549	35.22767	105.50330	16.626837	124.1041	98
Oct 2017	87.56754	52.38011	122.75497	33.753007	141.3821	84
Nov 2017	82.45709	47.22011	117.69407	28.566785	136.3474	75
Dec 2017	89.87369	54.58724	125.16015	35.907717	143.8397	81
Jan 2018	106.74014	71.40440	142.07588	52.698787	160.7815	66
Feb 2018	83.95034	48.56526	119.33542	29.833529	138.0671	81
Mar 2018	67.19598	31.76163	102.63033	13.003816	121.3881	119
Apr 2018	68.44031	32.95676	103.92386	14.172907	122.7077	105
May 2018	73.11653	37.58384	108.64921	18.773978	127.4591	100
Jun 2018	63.24297	27.66122	98.82472	8.825382	117.6606	123

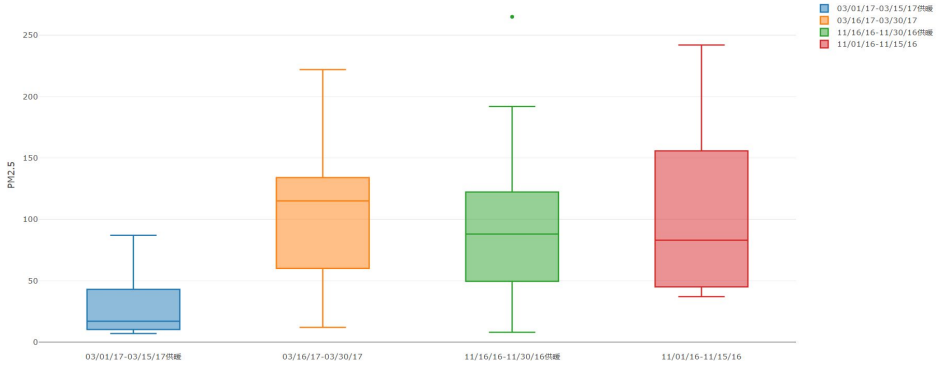
Seasonal Patterns of PM 2.5 Concentration

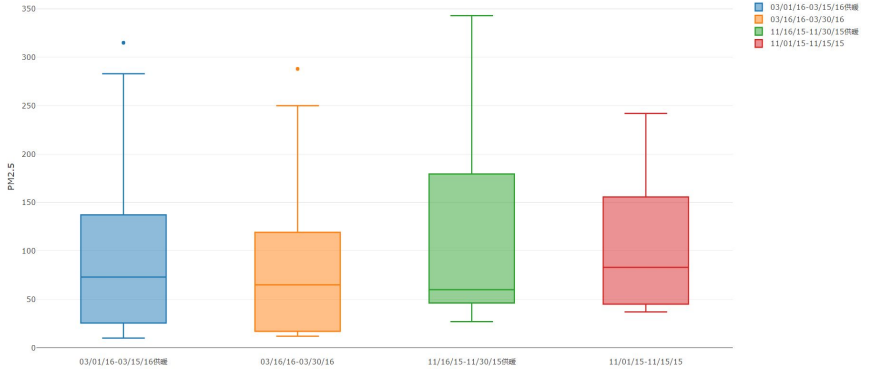
Does the heating in winter influence the pm 2.5 values?

Every year heating in winter begins at Nov 15 and ends in Mar 15 next year.

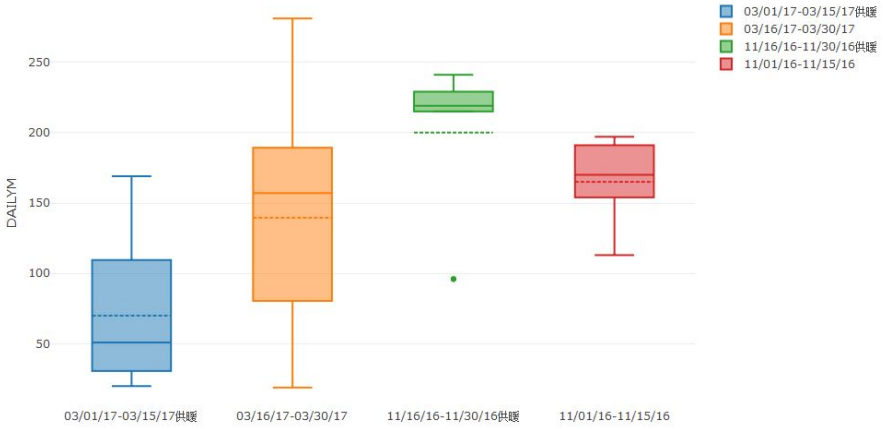
Comparison for daily average pm2.5 values

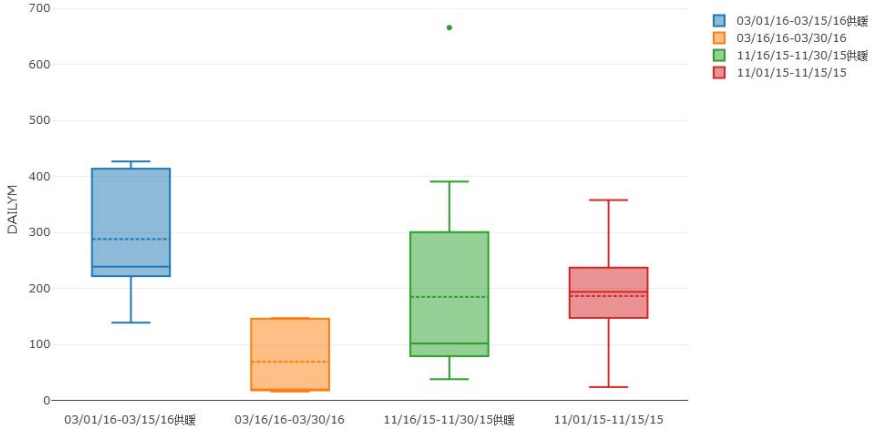


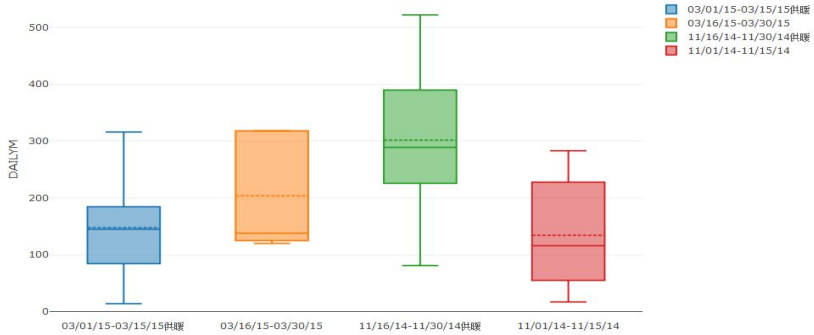




Comparison for daily maximum pm2.5 values







APEC Blue

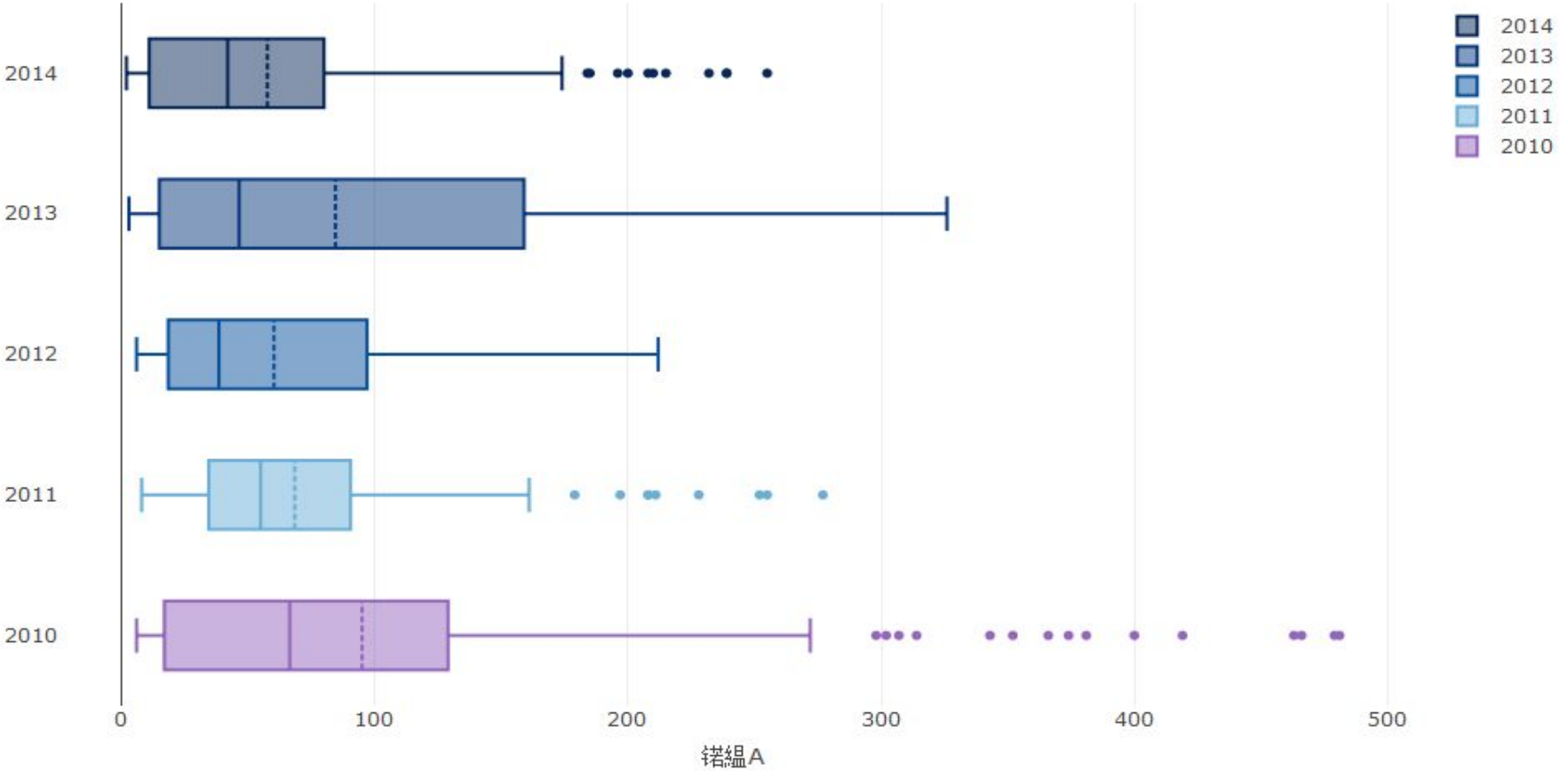
the rare blue sky in Beijing during APEC CHINA 2014 due to emission reduction campaign directed by China government

Are these policies issued by China government effective?

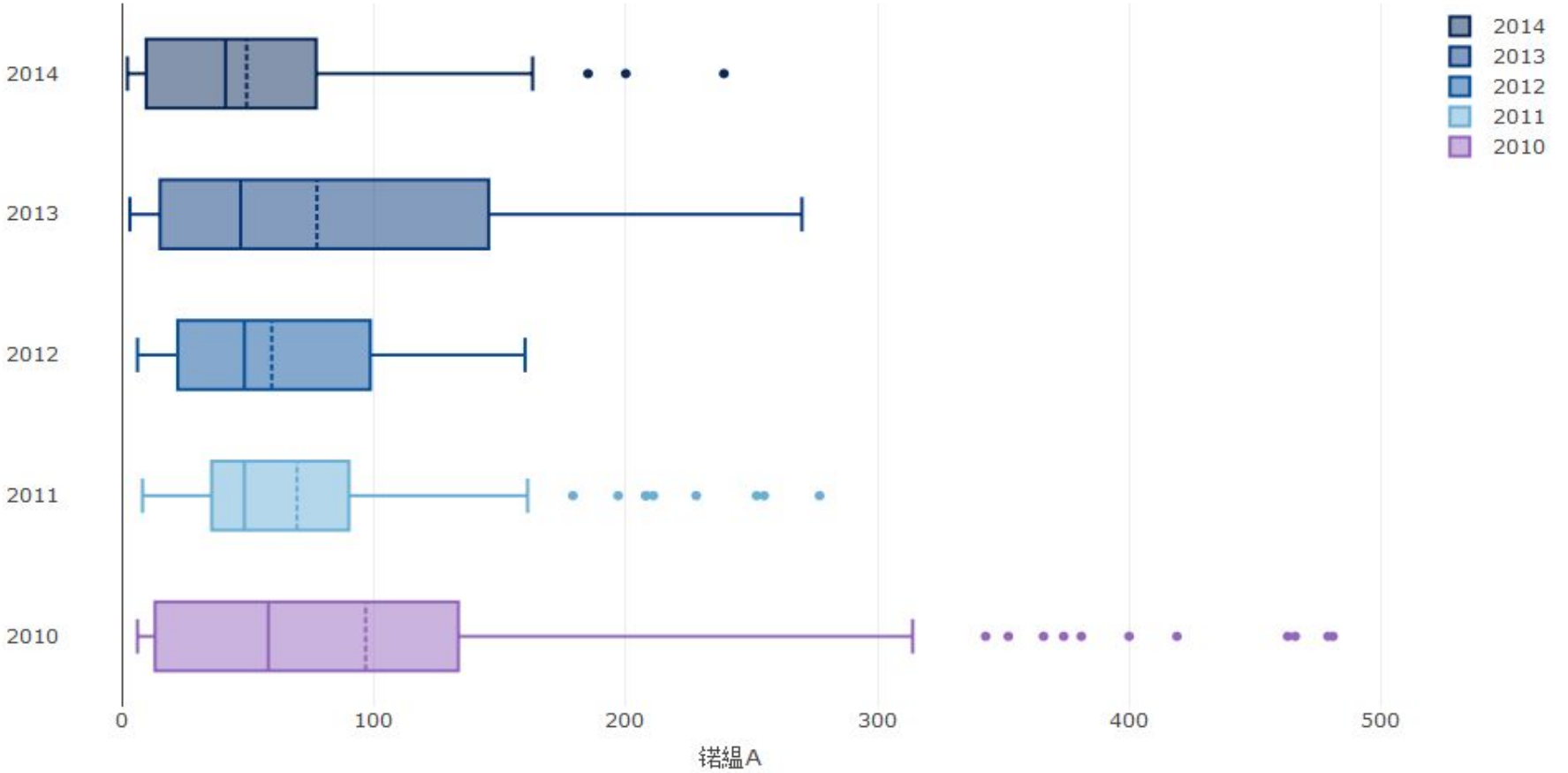
Two stages of strict measures on controlling air pollution:

1. 2014/11/03 -- 2014/11/05
2. 2014/11/06 -- 2014/11/12

11/03 -- 11/12



11/06 -- 11/12


2014
2013
2012
2011
2010
2014
2013
2012
2011
2010
0
100
200
300
400
500
锘緰A

Results of two-sample t-test

	p-value
2014 vs 2013	2.559e-05
2014 vs 2012	0.03346
2014 vs 2011	0.0001492
2014 vs 2010	6.642e-07

Diurnal patterns of PM 2.5 Concentration

Hourly data from U.S. Department of State

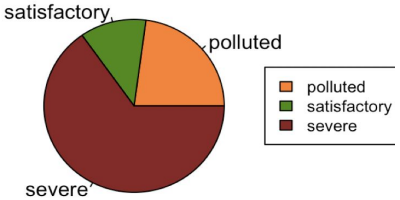
1. Severe days distribution
2. Hourly average in different seasons
3. Weekend and Weekday

Severe Days calculated by Daily Maximum

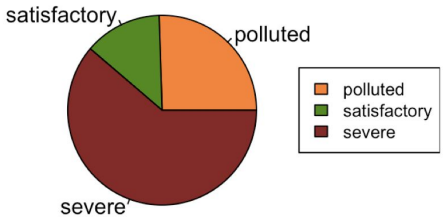
Severe Days calculated by Daily max					
	Satisfactory	Polluted	Severe	NA	Sum
2010	38	73	208	45	364
2011	39	75	180	71	365
2012	43	67	192	64	366
2013	32	85	213	35	365
2014	40	85	208	32	365
2015	71	105	162	27	365
2016	85	114	143	24	366
2017	51	52	60	18	181

Satisfactory	pm2.5<60
Polluted	60≤pm2.5<120
Severe	pm2.5≥120

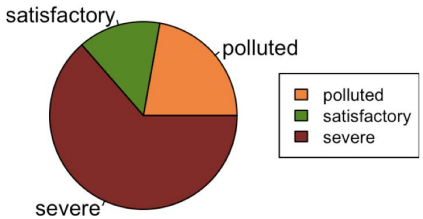
PIE CHART OF 2010



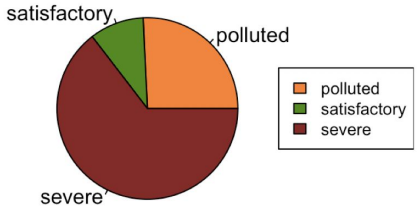
PIE CHART OF 2011



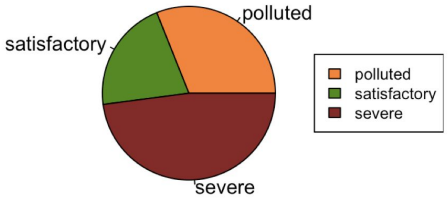
PIE CHART OF 2012



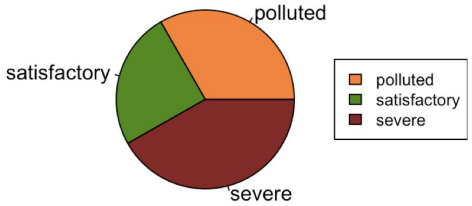
PIE CHART OF 2013



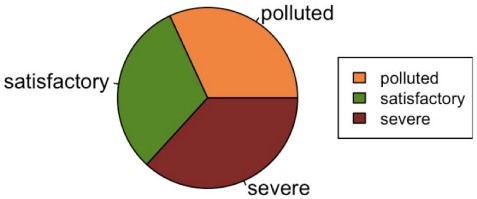
PIE CHART OF 2015



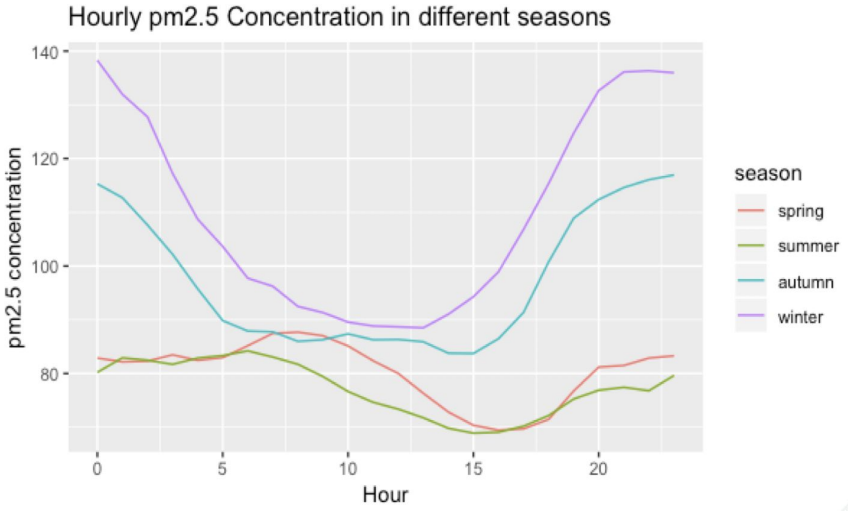
PIE CHART OF 2016



PIE CHART OF 2017

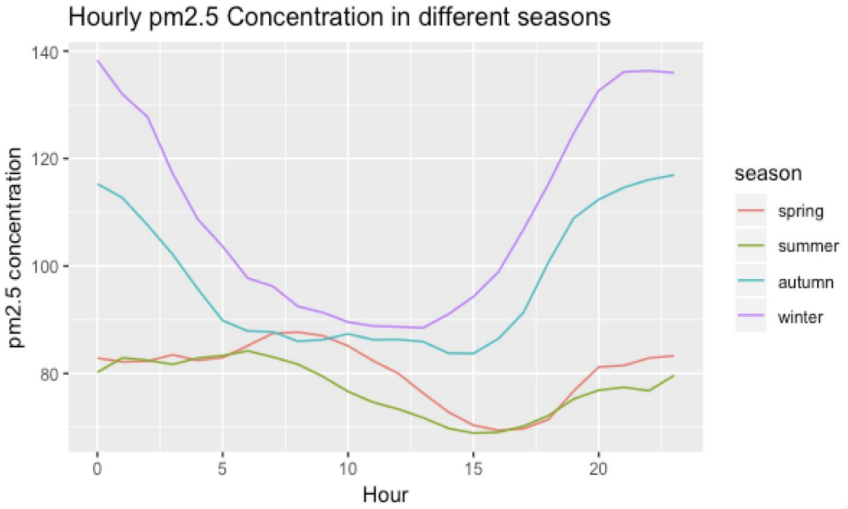


Hourly Average in Different Seasons



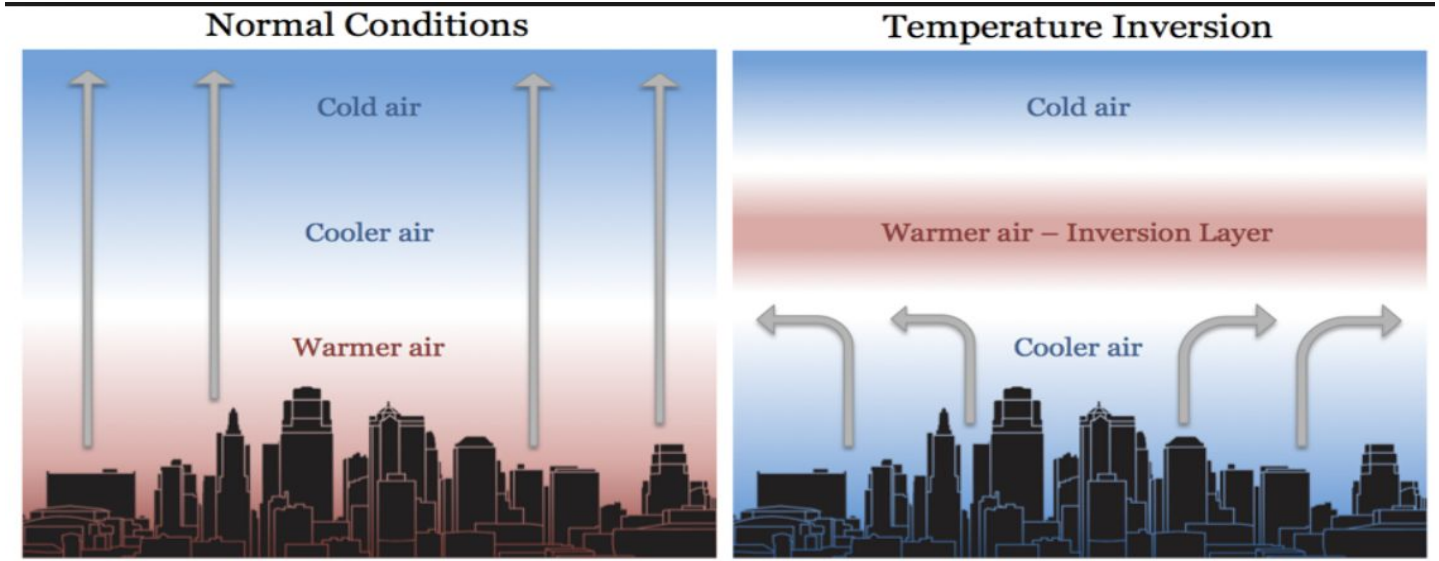
- ➔ Seasons : Mar-May Spring; Jun-Aug Summer; Sep-Nov Autumn; Dec-Feb Winter.
- ➔ Group different seasons from 2010.1.2-2017.6.30, calculate hourly average.

Results



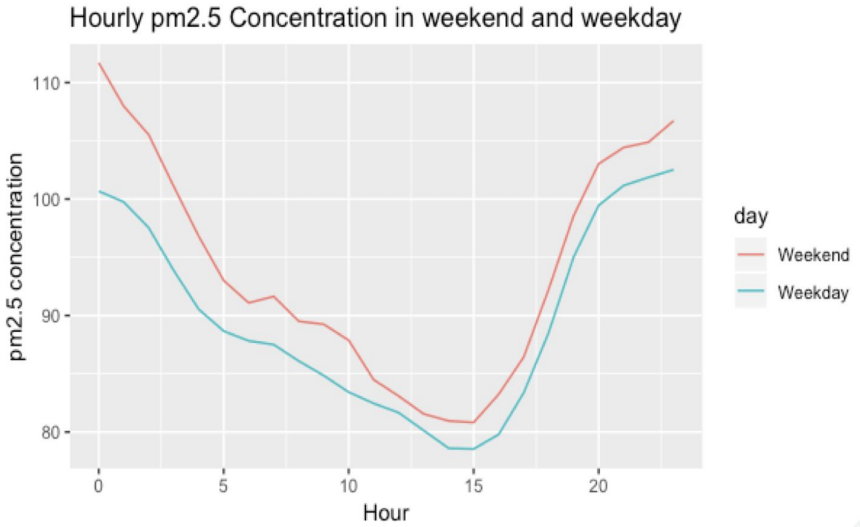
1. The daily pattern presents systematic seasonal variations.
2. Mobile-source influence:The peak rises at 18:00 p.m.-20:00 p.m., this indicates the evening rush hour peak. In summer and autumn, the peak rises at 6:00 a.m.-10:00 a.m., this indicates the morning rush hour peak.
3. The pm2.5 still stays high in the evening. This is because temperature inversion.

Temperature Inversion—Pollution Trapped



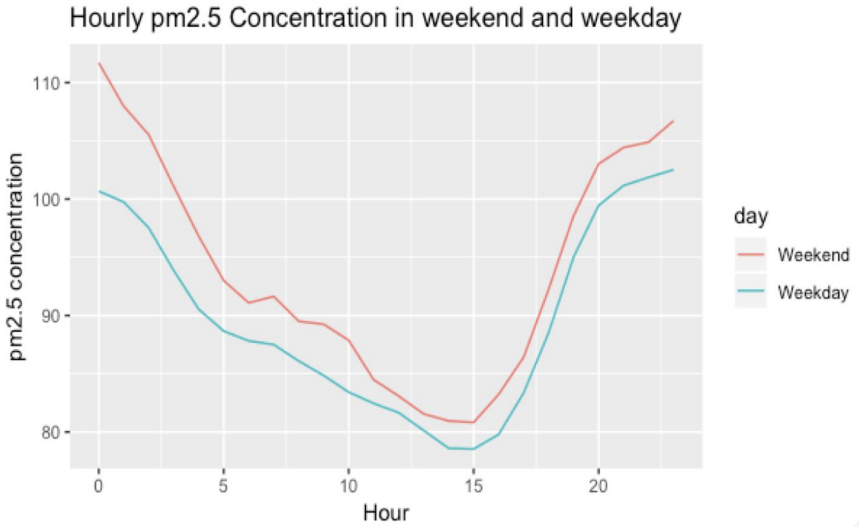
Arrows show air flow in normal conditions on the left and during temperature inversion on the right. In normal conditions, warm air rises and normal convective patterns persist. During temperature inversion, the warm air acts as a cap, effectively shutting down convection and trapping smog over the city.

Day of Week Patterns



Group weekends (Saturdays and Sundays) and weekdays from 2010.1.2-2017.6.30, and take hourly average.

Results



1. The yearly average PM2.5 concentration is higher on weekend. This is probably due to the driving restriction in Beijing, that is, about 20% of cars have to stay off the road on each weekday.

2. Since the vehicle possession level amounts to close to 5.35 million units*, this means there will be 1.07 million more cars on the road on weekends than weekdays.

*Data source: http://calendar.hexun.com/area/dqzb_110000_D0940000.shtml

Pollution Source

- Vehicles
- Coal-burning
- Airborne dust
- Mobile sources
- Industrial production

Suggestion

- Heating: ① It's far more efficient to heat into a system that can warm an entire city than to heat buildings individually with boilers. ② heating in winter should be timely adjusted according to actual temperature change.
- Reduce car dependence and to ban barbecuing and straw-burning
- Encourage an economic transition toward clean fuels to lessen the need of coal and convert waste products into fuel by applying new technologies
- In government administration, harsher supervision and enforcement of penalization should be applied