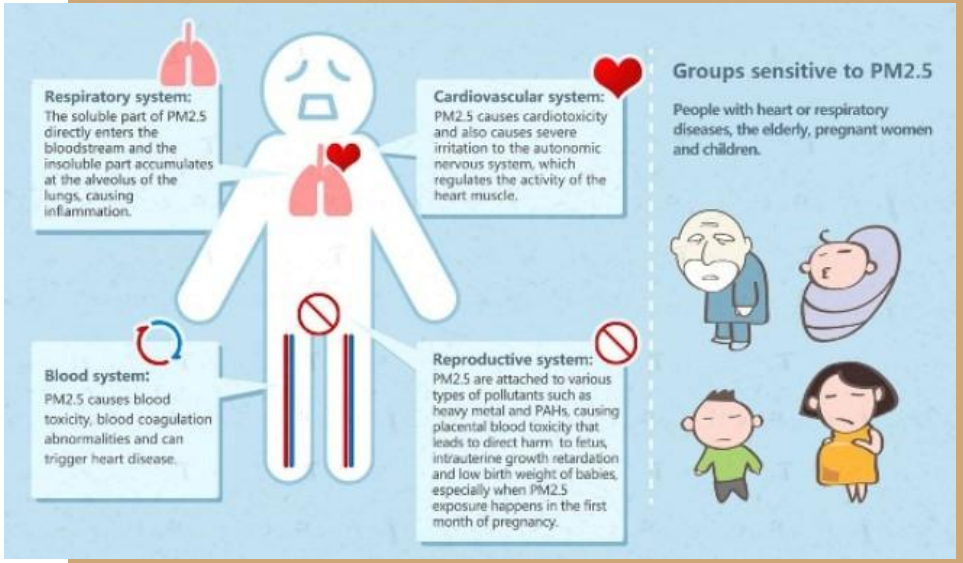
Beijing : PM2.5

Xinning Chu, Dandan Feng,
Kang Fu, Ashley Hall

PM2.5

- Particulate matter enters the body through the nose and mouth
- The body will eliminate larger particles
- PM10 - less than 10 microns in diameter
  - Dust, pollen, mold, etc
- PM2.5 - dangerous particles of pollutants that are less than 2.5 microns in diameter
  - Combustion particles, organic compounds, metals, etc.



THE BIGGER PICTURE

Air pollution

India is home to six of ten cities in the world with the worst air pollution, according to the World Health Organisation.

Most polluted **Least polluted cities and towns**
Data collected in 4,357 settlements* from 2010 to 2016

MOST POLLUTED
PM 2.5** annual mean, micrograms per m^3

City		Country
Gwalior	176	India
Allahabad	170	India
Al Jubail	152	Saudi Arabia
Pasakha	150	Bhutan
Raipur	144	India
Novi Sad	142	Serbia
Delhi	123	India
Ludhiana	122	India
Cairo	117	Egypt
Khanna	114	India

LEAST POLLUTED
PM 2.5 annual mean, micrograms per m^3

City/town		Country
Bredkalen	1	Sweden
Muonio	2	Finland
Dias D'Avila	3	Brazil
El Pueyo De Araguas	3	Spain
Guimaraes	3	Portugal
La Plaine Des Cafres	3	France
Lousame	3	Spain
Kiruna	3	Sweden
Te Anau	3	New Zealand
Arrest	4	France

Note: Data is a combination of measured and converted values.
*Ranging in size from under 100 people to more than 10 million inhabitants.
**The concentration of fine suspended particles of less than 2.5 microns in diameters is a common measure of air pollution.
Source: World Health Organisation.

C. Hughes, 02/05/2017 REUTERS

September 2013 - State Council unveiled the “Action Plan for the Prevention and Control of Air Pollution” - Aimed to increase air quality over 5 year period

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

MARCH 2014 - US Embassy in Beijing releases a trove of air quality data

“Fuels a big step at the government level: declaring war on air pollution. One of the reasons for that is that the health argument was very strongly presented, and the fact that the citizens were really breathing air that was totally unacceptable.”

Inspired Chinese officials to launch their own air monitoring operations around the country

1. What does this Air Quality Data look like?

April
2008

June
2017

Microgram values observed hourly for each month in each year

Site	Parameter	Date (LST)	Year	Month	Day	Hour	Value	Unit	Duration	QC Name
Beijing	PM2.5	4/8/2008 15:00	2008	4	8	15	207	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 16:00	2008	4	8	16	180	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 17:00	2008	4	8	17	152	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 18:00	2008	4	8	18	162	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 19:00	2008	4	8	19	171	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 20:00	2008	4	8	20	219	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 21:00	2008	4	8	21	86	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 22:00	2008	4	8	22	63	μg/mg³	1 Hr	Valid
Beijing	PM2.5	4/8/2008 23:00	2008	4	8	23	61	μg/mg³	1 Hr	Valid

TO KEEP IN MIND WHILE SIFTING THROUGH DATA:

The US converts PM2.5 observations into an air quality index where readings at 50 or below are considered good quality

AIR QUALITY GOOD
0 - 50 INDEX

AIR QUALITY MODERATE
51 - 100 INDEX

AIR QUALITY UNHEALTHY
FOR SENSITIVE GROUPS
101 - 200 INDEX

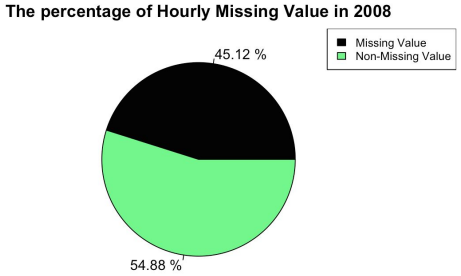
AIR QUALITY UNHEALTHY
151 - 200 INDEX

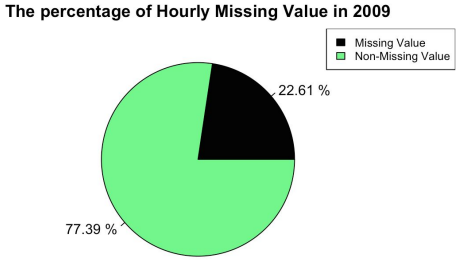
Project objective

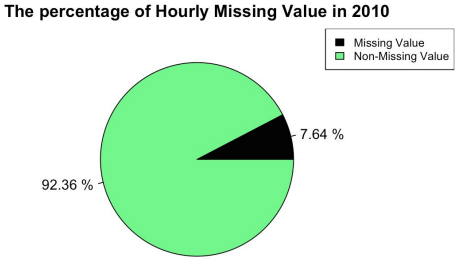
Using historical PM2.5 values, is Beijing’s air quality increasing or decreasing?

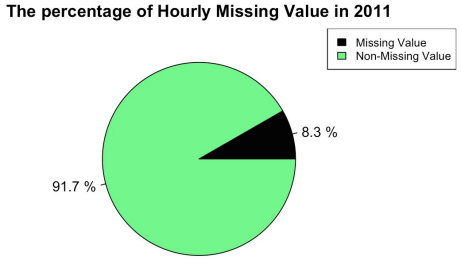
Some Initial Exploratory Analysis of Beijing's air quality data released by its US Embassy

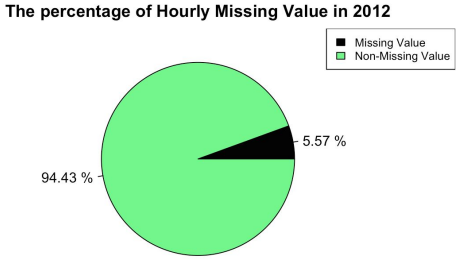
1. Missing values in the hourly data sets

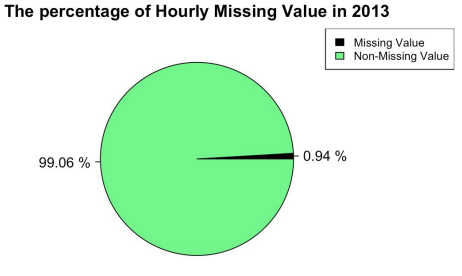




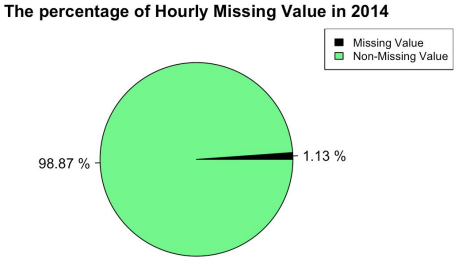


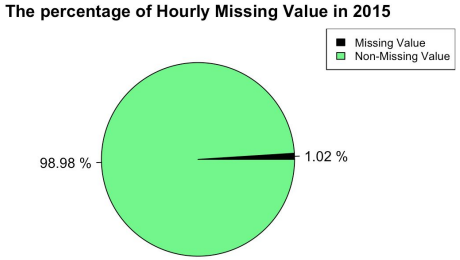


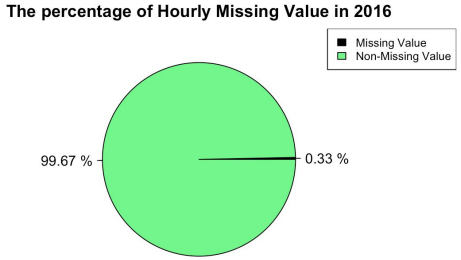




1. Missing values in the hourly data sets







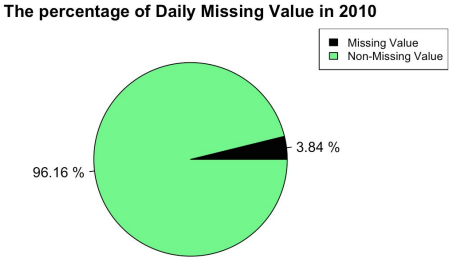
For the ratios of missing values in 2008 and 2009 are too high, which will definitely affect the accuracy of the future data analysis. We decided not to use the data from 2008 to 2009, only focusing on the data from 2010 to 2016.

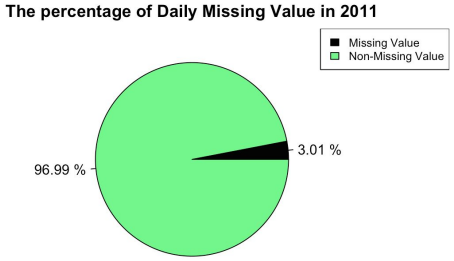
2. Data preprocessing

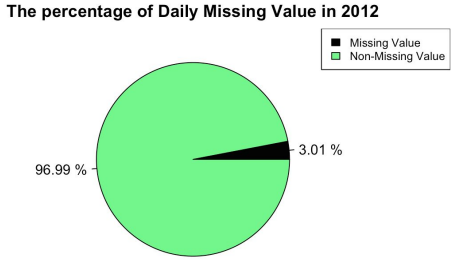
For the hourly data can be easily affected by many external factors, using the hourly data directly to do data analysis is not very reasonable. So, after discussing in group, we decided to do the follow data preprocessing:

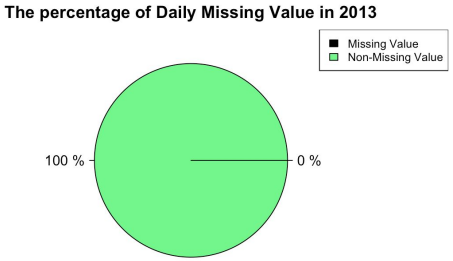
1. Delete the observations that have missing PM2.5 values in the 2010 to 2016 data set.
2. Calculate the daily maximum PM2.5 values from 2010 to 2016 using the hourly data.

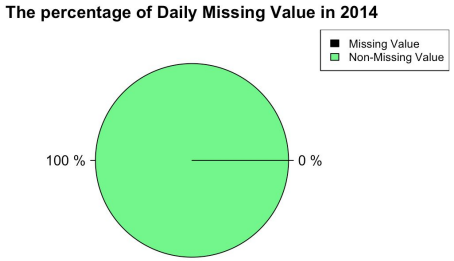
3. Missing values in the daily data sets

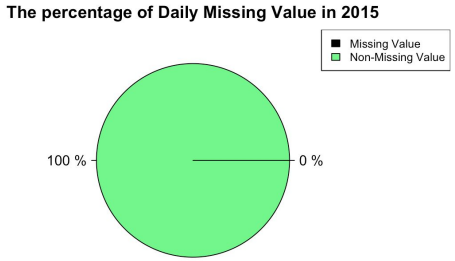




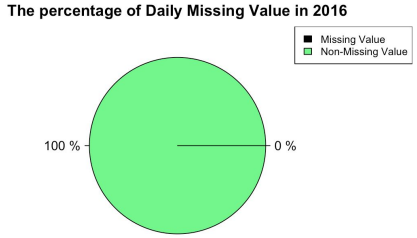






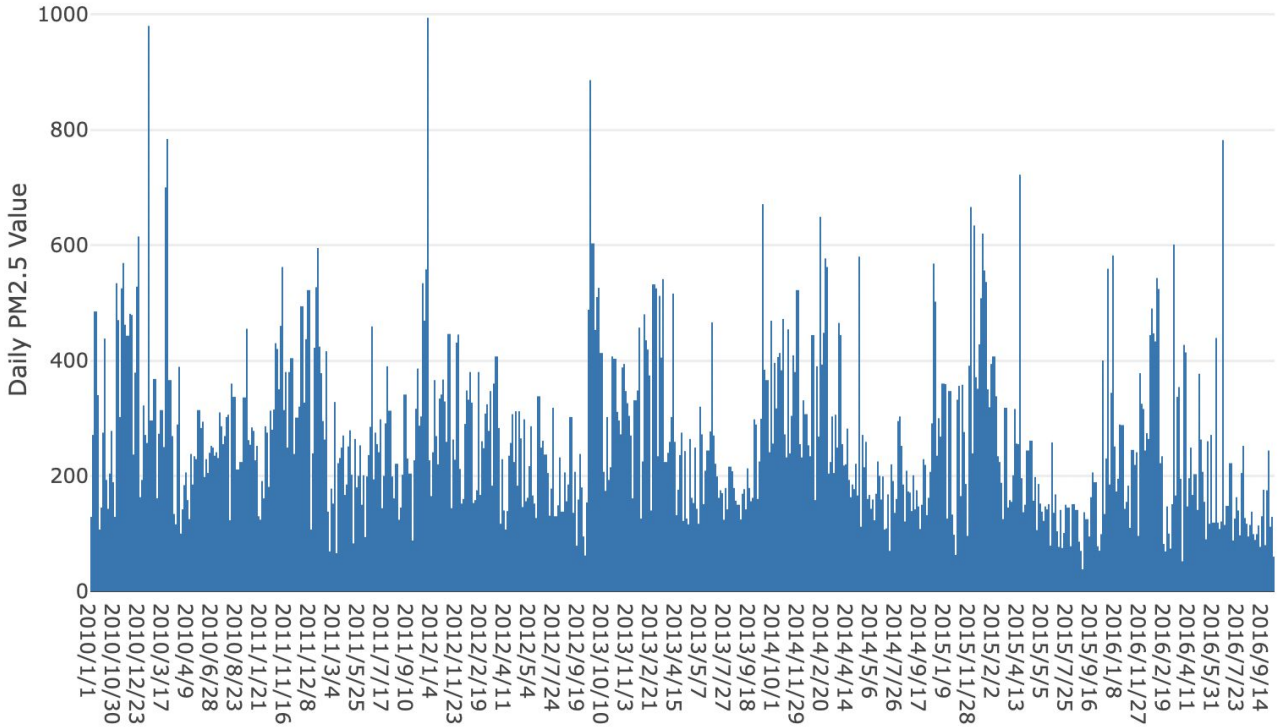


3. Missing values in the daily data sets



From these pie charts we can see that the calculated daily maximum PM2.5 data is not only more stable and more reasonable to be used in data analysis, but also have a lower ratio of missing values.

4. Barplot for the daily maximum PM2.5 values V.S. date

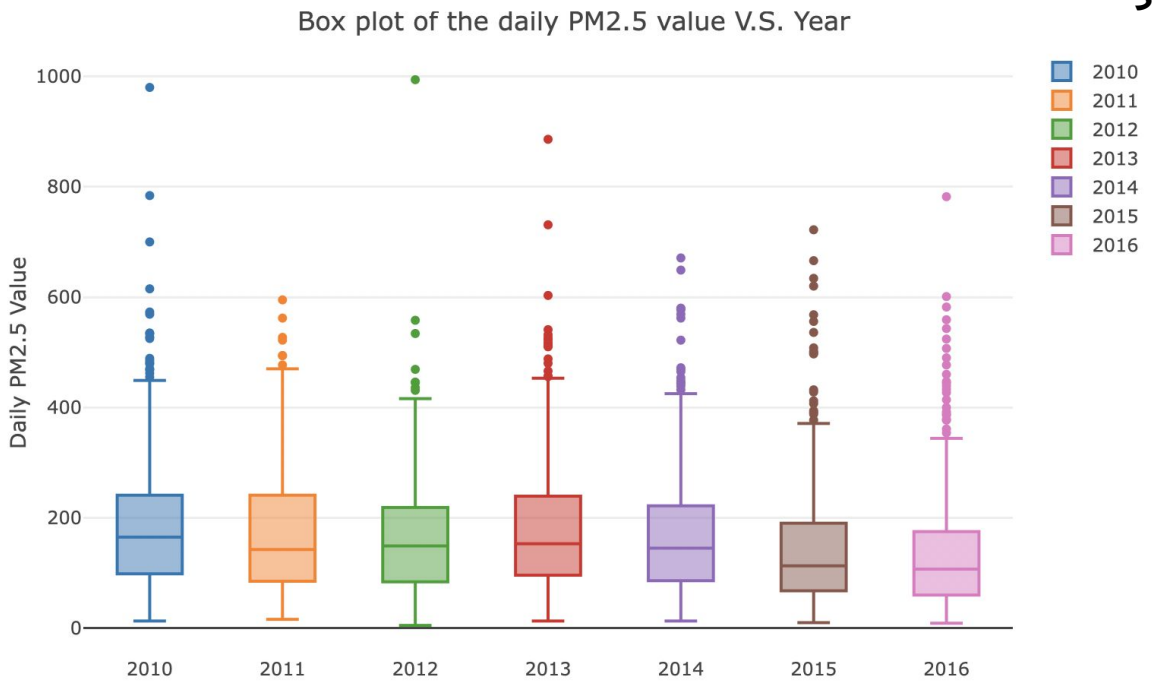


5. Information from the Barplot of the daily data

1. The PM2.5 value does is time-varying.
2. This kind of time-dependent changes is more likely a periodical pattern (seasonal pattern).
3. Overall, the PM2.5 value seems to be stationary from 2010 to 2016.

6. More plots derived from the daily data

Stationary?



7. Potential statistical method

1. Unpaired two sample t-test?

What is unpaired 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. The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if X1, X2, ..., Xn and Y1, Y2, ... , Yn are the two samples, then Xi corresponds to Yi. For paired samples, the difference Xi - Yi is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.

However,

✖ Note that, unpaired two-samples t-test can be used only under certain conditions:

- when the two groups of samples (A and B), being compared, are **normally distributed**. This can be checked using **Shapiro-Wilk test**.
- and when the **variances** of the two groups are equal. This can be checked using **F-test**.

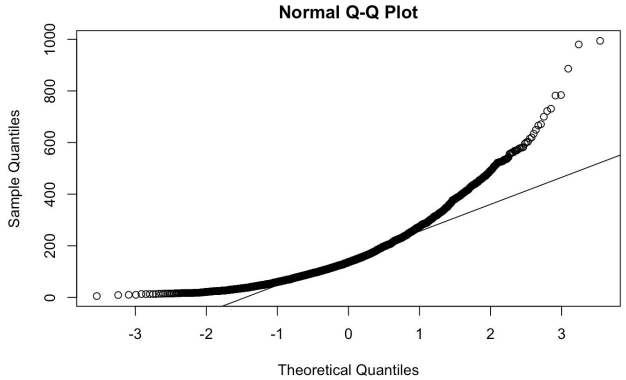
8. Check normality assumption for daily PM2.5 value

Methods:

1. Q-Q plot:
   A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions.
2. KS-test:
   In statistics, the Kolmogorov–Smirnov test (K–S test or KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). It is named after Andrey Kolmogorov and Nikolai Smirnov.

9. Results of normality check

1. Q-Q plot:



Normality assumption doesn't hold, NO two sample t-test!!!!!!

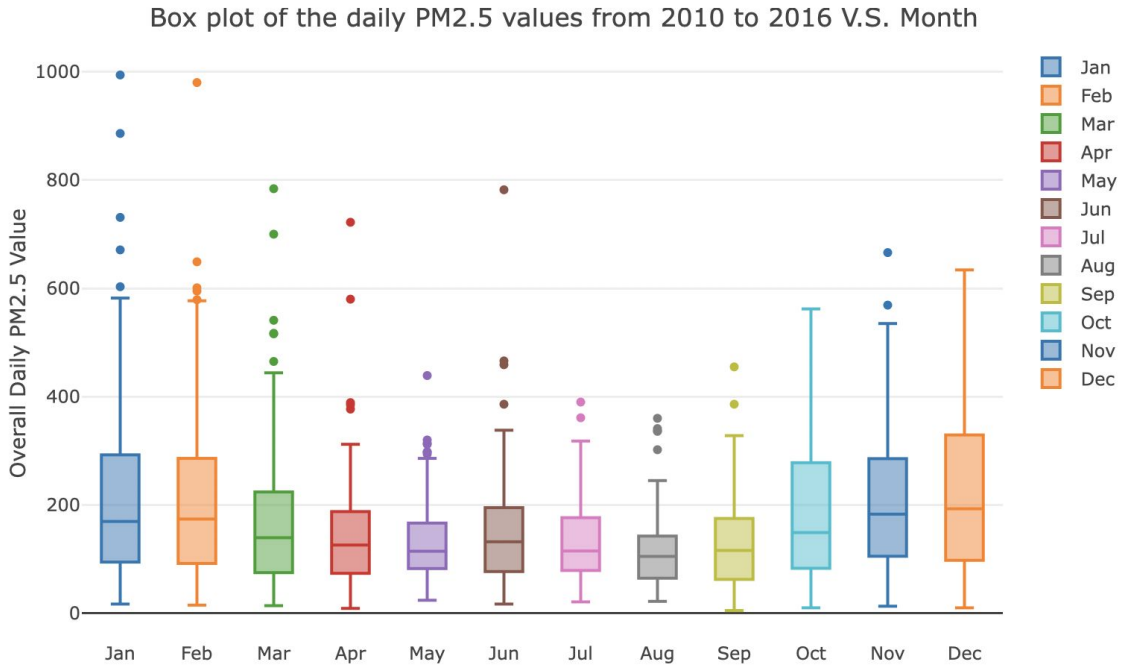
2. KS-test:

```
One-sample Kolmogorov-Smirnov test

data:  Dailyall$PM2.5Value
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

10. More valid statistical method

Time series analysis!

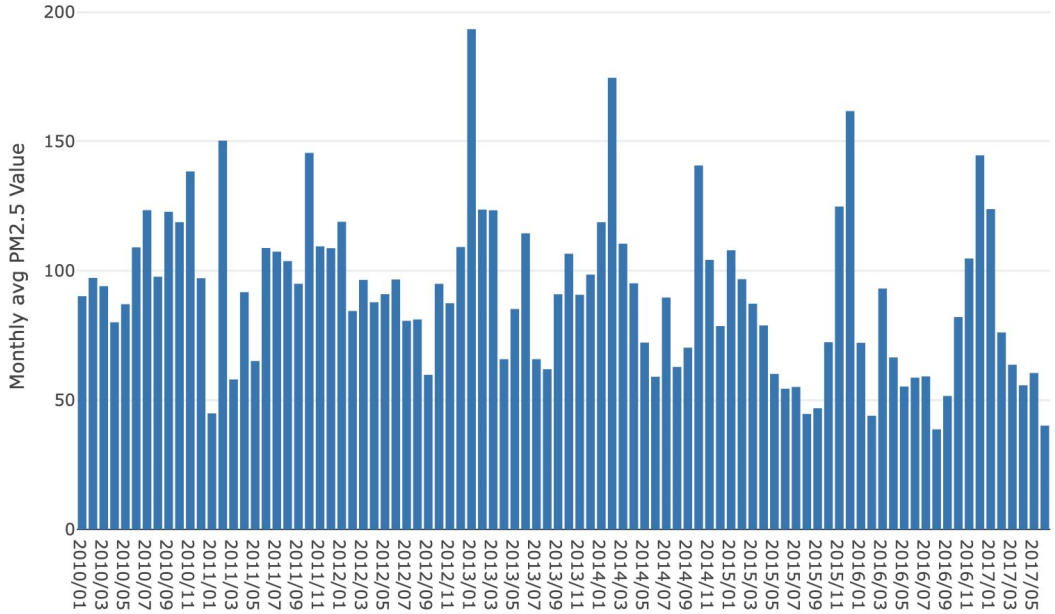
1. The data itself is in time series, a series of data points indexed (or listed or graphed) in time order.



2. Seasonal pattern

Preparation to the time series analysis

Calculate monthly average data:



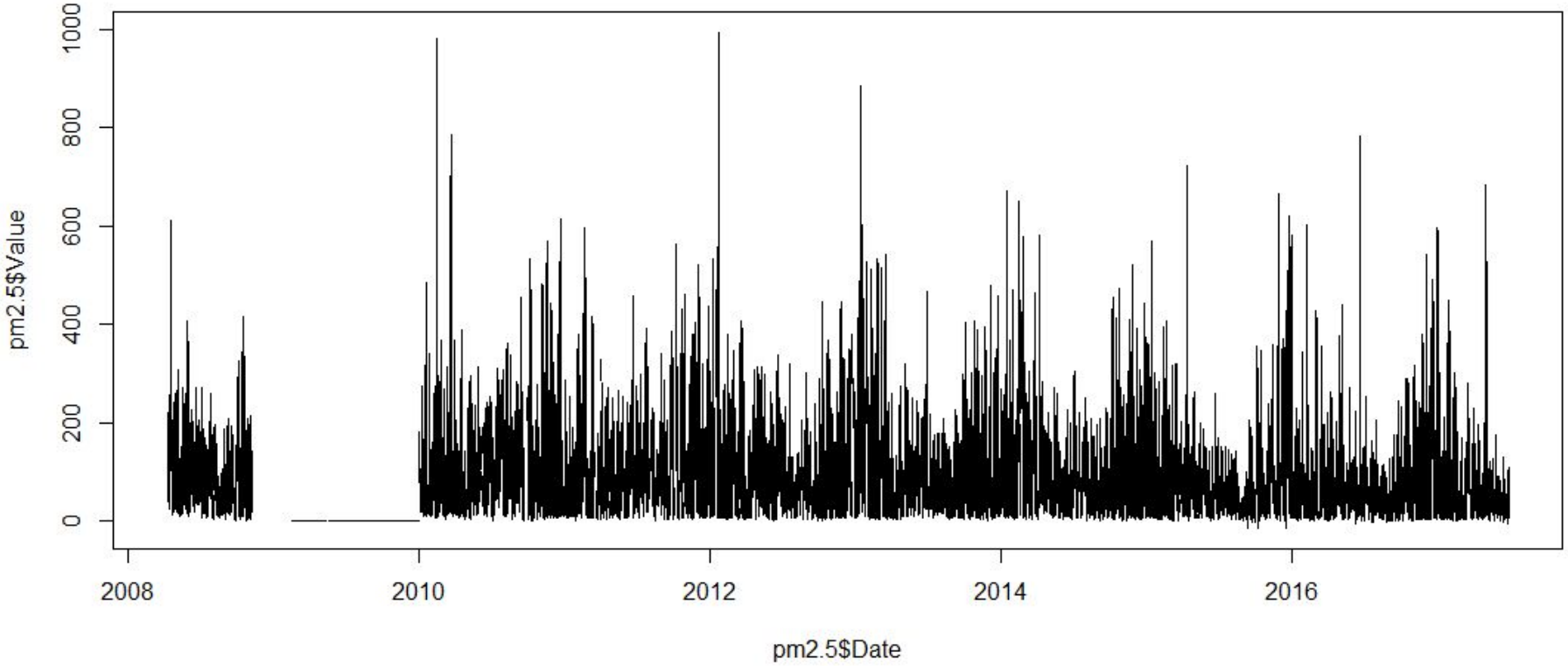
Time Series Analysis

Methods:

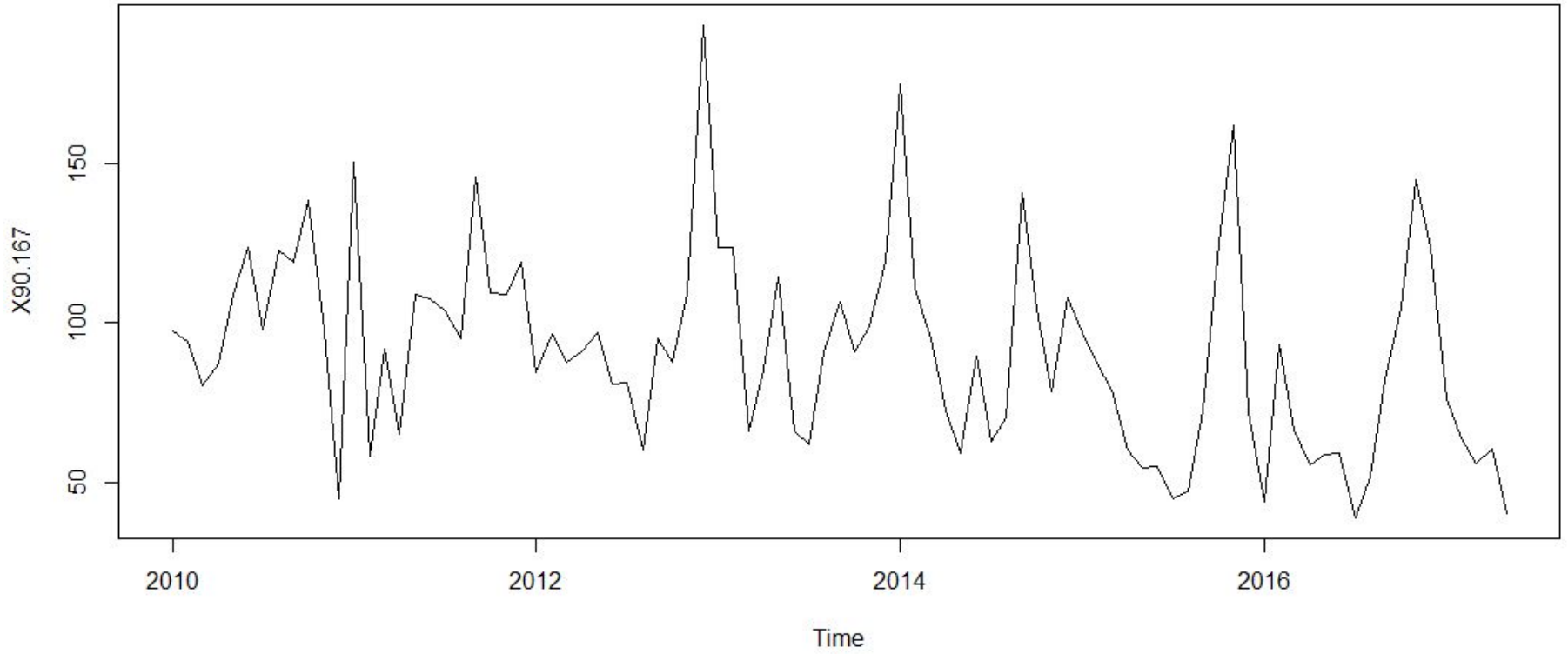
1. Making plots (time plot, ACF/PACF plots)
2. Stationary tests: Augmented Dickey-Fuller test; Phillips-Perron test
3. Build the ARIMA / SARIMA model
4. Diagnostics and forecasting

R packages used:
astsa / tseries / forecast / aTSA

Time plot of hourly average pm2.5 values from 2008 to 2017



Time plot of monthly average pm2.5 values from 2010 to 2017



The SARIMA model is given by

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t,$$

where wt is the usual Gaussian white noise process. The general model is denoted as ARIMA(p, d, q) × (P, D, Q)s.

The pure seasonal autoregressive moving average model, ARMA(P,Q)s takes the form

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t,$$

where the operators

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

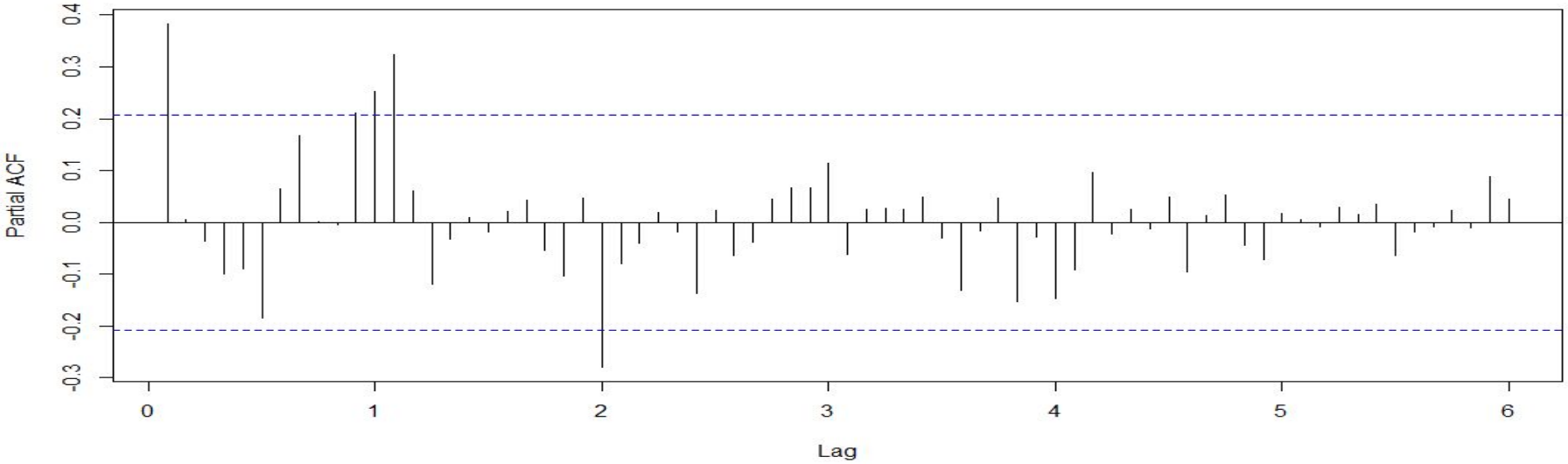
and

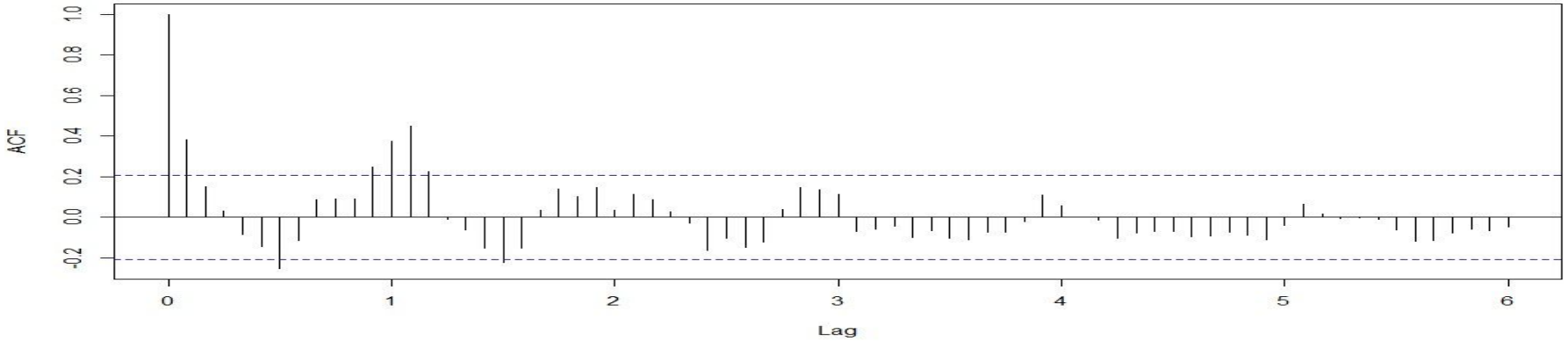
$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$$

Behavior of the ACF and PACF for Pure SARIMA Models

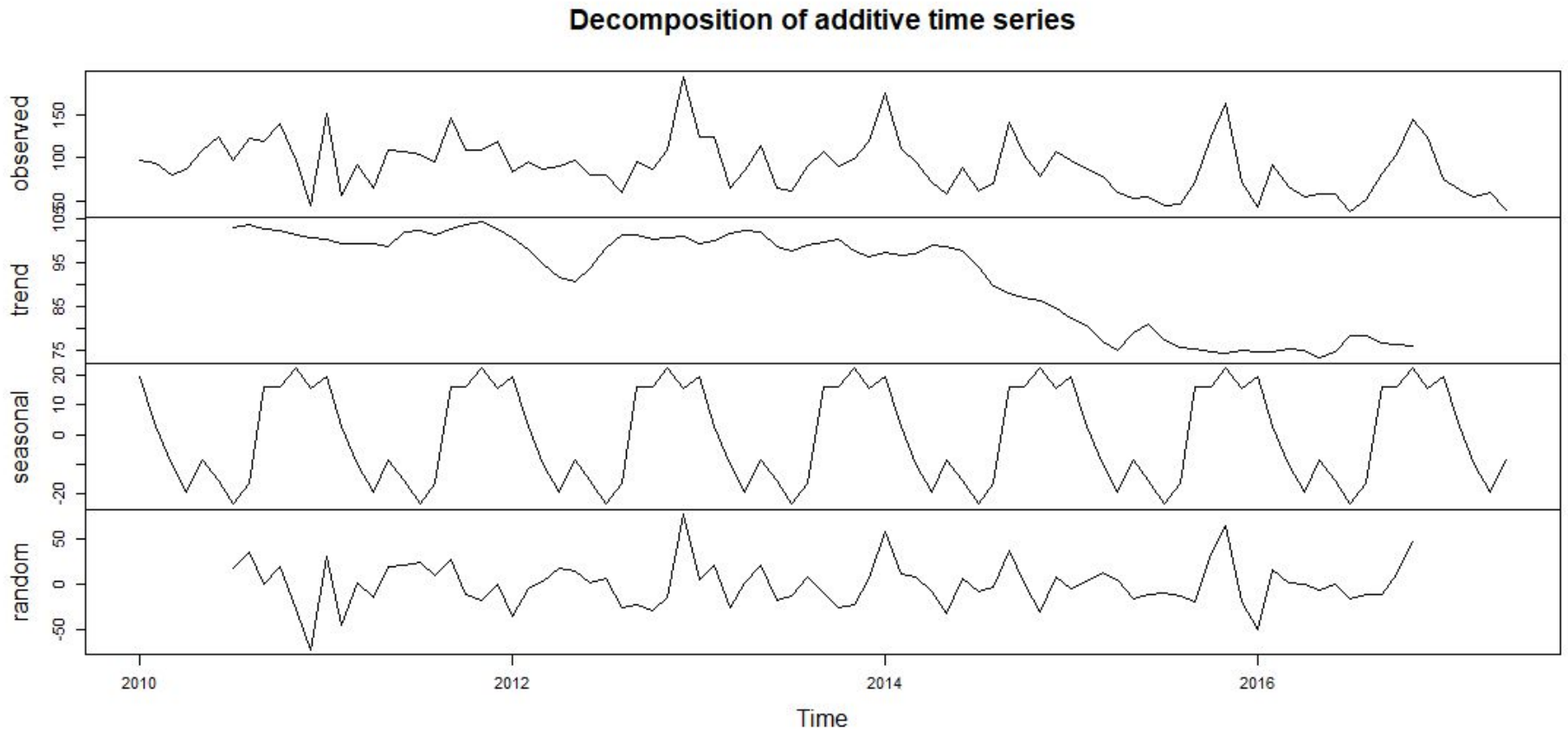
	$\text{AR}(P)_s$	$\text{MA}(Q)_s$	$\text{ARMA}(P, Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots,$ are zero.

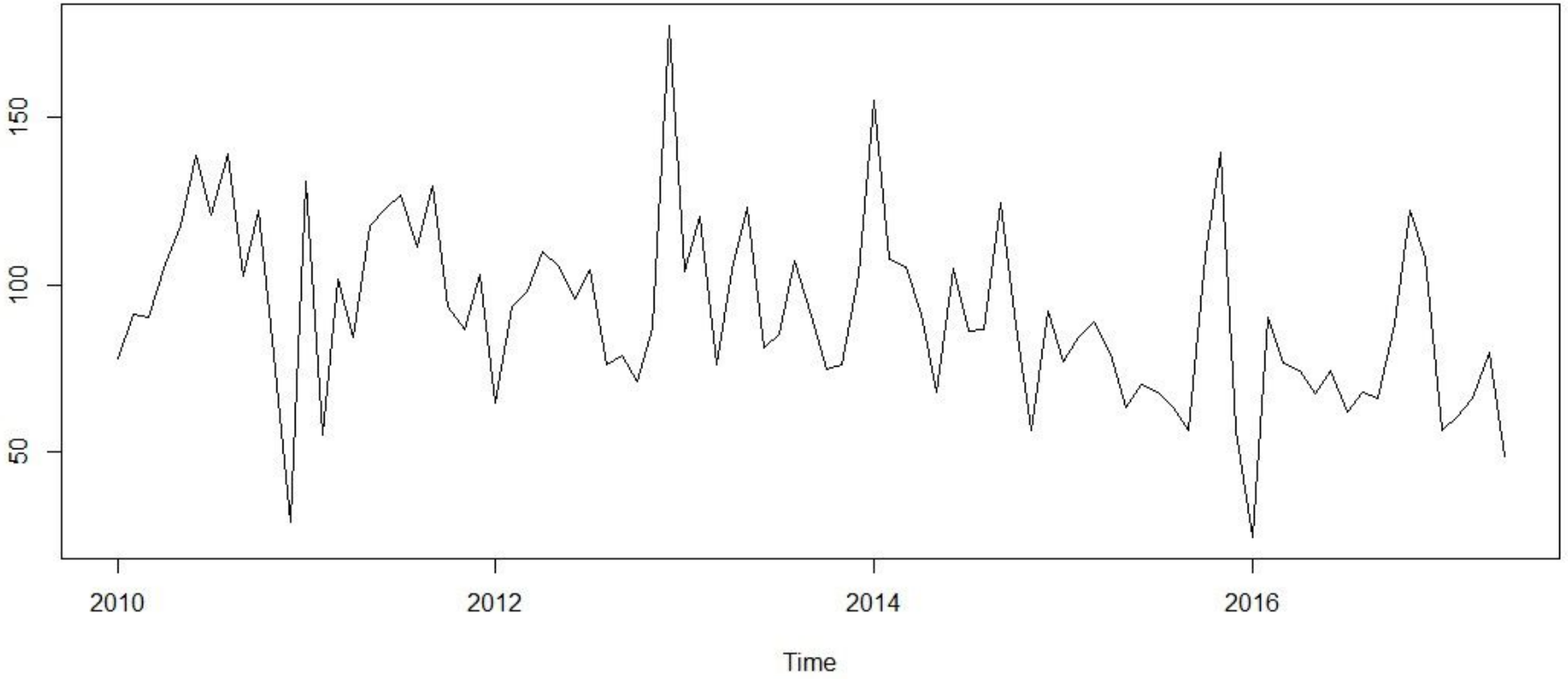

Partial ACF
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0
1
2
3
4
5
6
Lag


ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
1
2
3
4
5
6
Lag

Time plot of pm2.5 values without seasonal factor



Time plot of pm2.5 values without seasonal factor



Augmented Dickey-Fuller test (adf test) :

1. The null hypothesis : a unit root is present in the time series sample.
2. It is applied to the model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where α is a constant, β the coefficient on a time trend and p the lag order of the autoregressive process.

3. The unit root test is then carried out under the null hypothesis γ=0 against the alternative hypothesis of γ<0. Once a value for the test statistic

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey–Fuller Test. If the test statistic is less (this test is non symmetrical so we do not consider an absolute value) than the (larger negative) critical value, then the null hypothesis of γ=0 is rejected and no unit root is present.

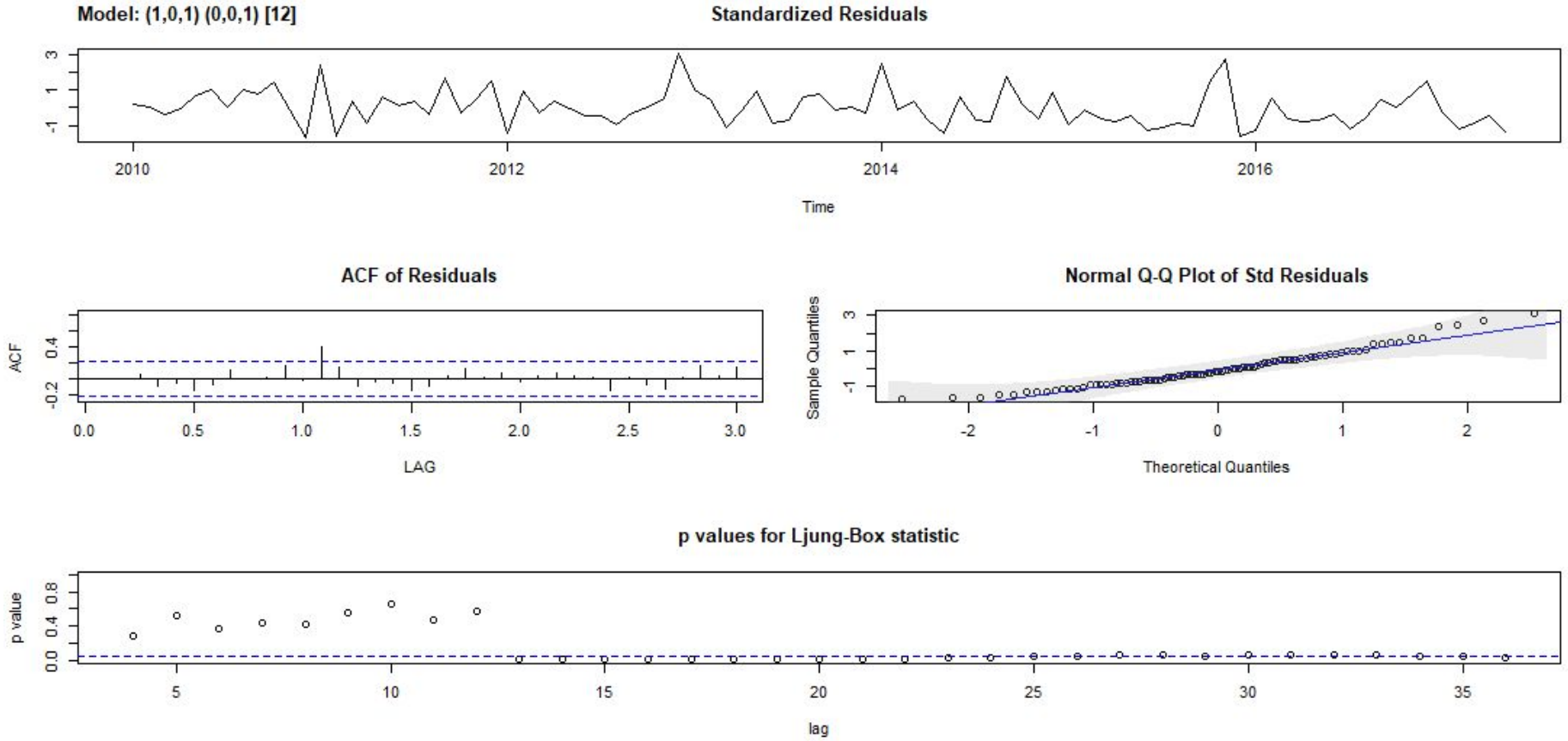
Results of stationary tests:

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

        Augmented Dickey-Fuller Test

data:  pm2.5noseason
Dickey-Fuller = -4.7766, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

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

        Phillips-Perron Unit Root Test

data:  pm2.5noseason
Dickey-Fuller Z(alpha) = -71.451, Truncation lag parameter = 3, p-value = 0.01
alternative hypothesis: stationary

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

Parameters Estimation:

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

```
Coefficients:
          ar1      ma1     sma1     xmean
      0.3607  -0.1591  0.4210  90.6653
s.e.  0.4002   0.4191  0.1387   5.1030


$ttable
      Estimate     SE t.value p.value
ar1     0.3607 0.4002  0.9013  0.3700
ma1    -0.1591 0.4191 -0.3796  0.7052
sma1    0.4210 0.1387  3.0350  0.0032
xmean  90.6653 5.1030 17.7671  0.0000

$AIC
[1] 7.669609

$AICc
[1] 7.700203

$BIC
[1] 6.781457
```

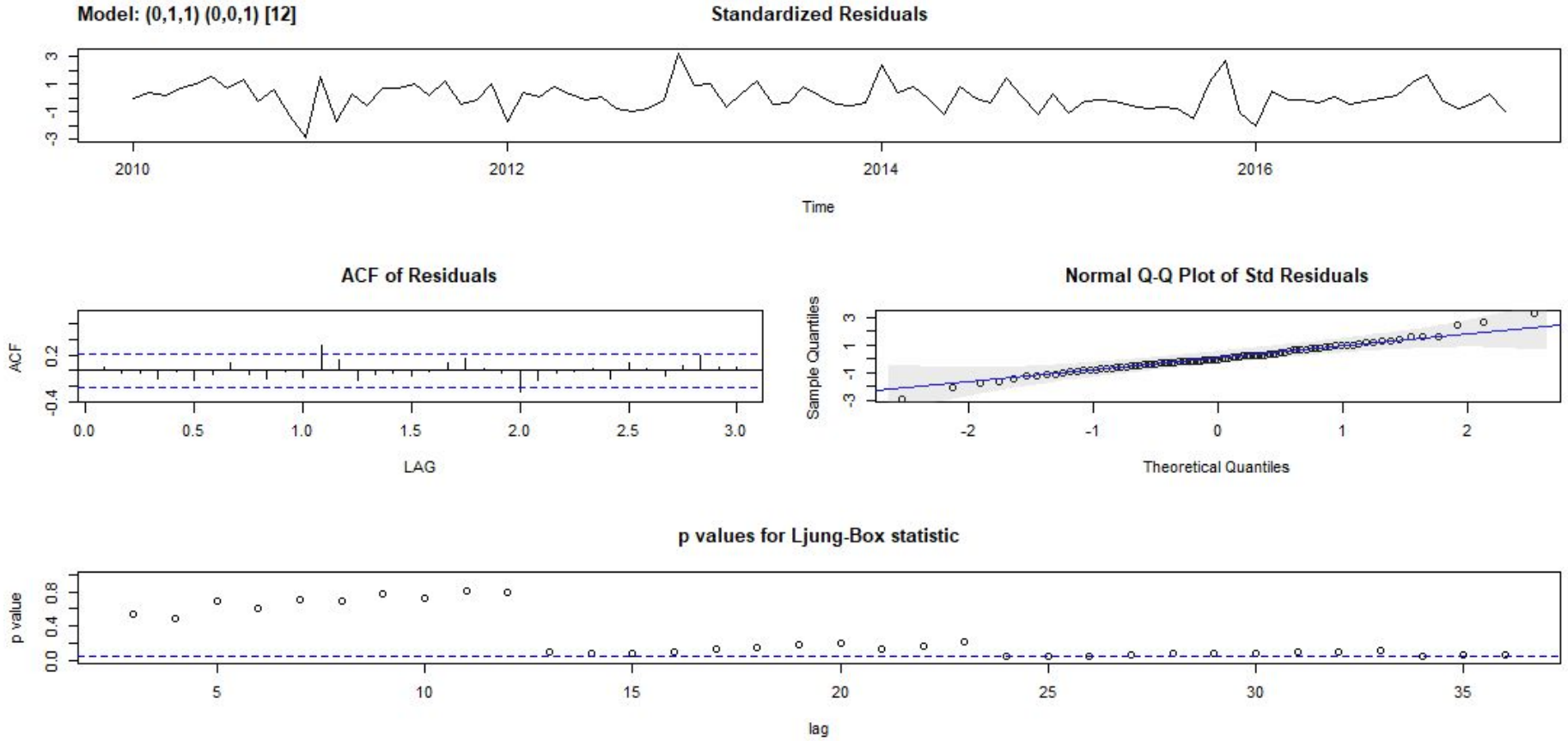

Model: (1,0,1) (0,0,1) [12]
Standardized Residuals
-1
1
3
2010
2012
2014
2016
Time
ACF of Residuals
ACF
-0.2
0.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-1
1
3
-2
-1
0
1
2
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
0.4
0.8
5
10
15
20
25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

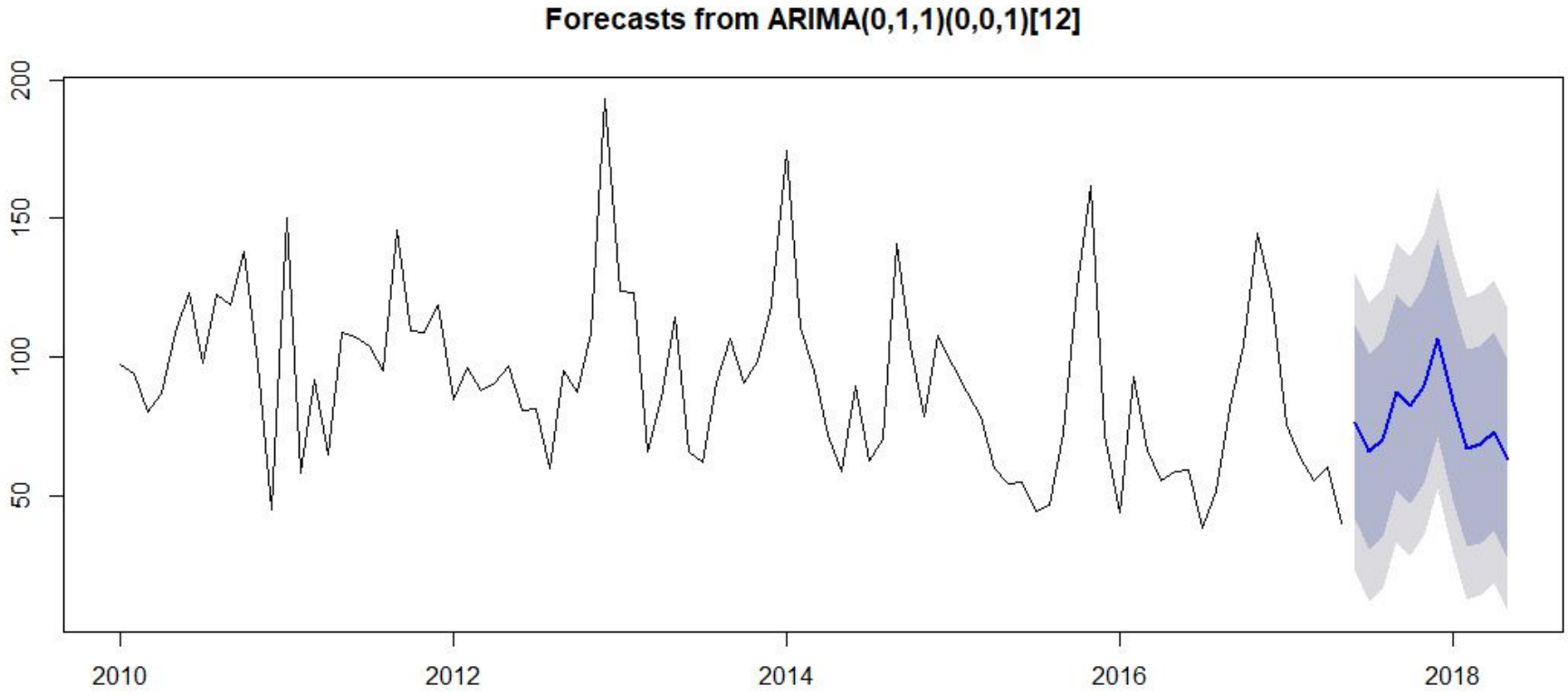
Parameters Estimation:

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

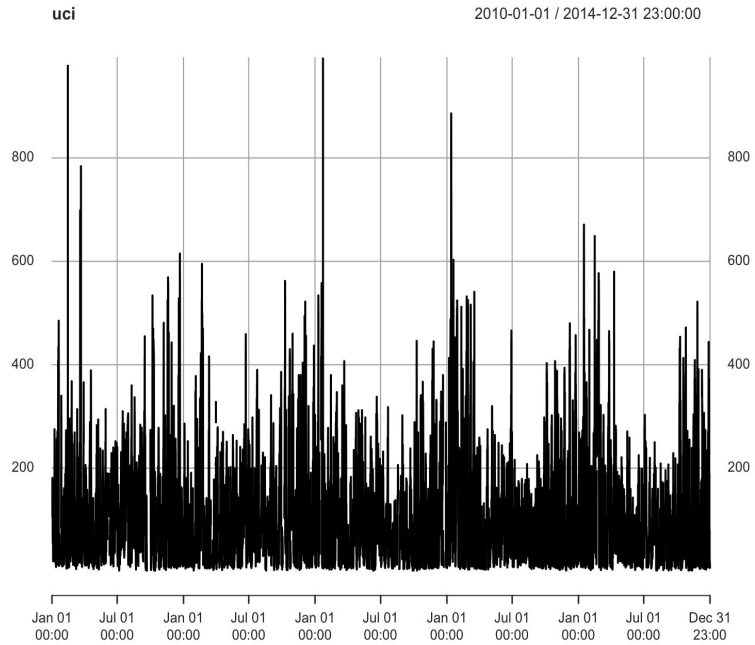

Hidden Factors Might Affect PM2.5

Data set from UCI website

- The dataset's time period is between Jan 1st, 2010 to Dec 31st, 2014 (43824 hourly observations and 8 variables)
- Attribute information:
  - Hourly PM2.5 concentration (ug/m^3)
  - Dew point
  - Temperature
  - Pressure (hPa)
  - Combined wind direction
  - Combined wind speed (m/s)
  - Cumulated hour of snow
  - Cumulated hours of rain

Summary Statistics & Time Series Plot

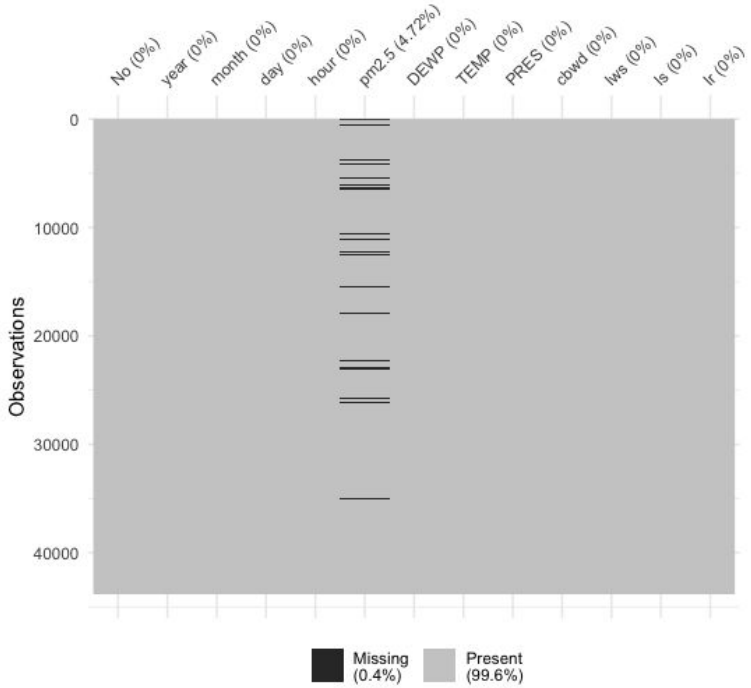
```
> summary(UCI)
       No              year          month             day              hour      
 Min.   :    1   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   : 0.00  
 1st Qu.:10957   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 5.75  
 Median :21912   Median :2012   Median : 7.000   Median :16.00   Median :11.50  
 Mean   :21912   Mean   :2012   Mean   : 6.524   Mean   :15.73   Mean   :11.50  
 3rd Qu.:32868   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:17.25  
 Max.   :43824   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.   :23.00  

     pm2.5             DEWP             TEMP             PRES           cbwd      
 Min.   :  0.00   Min.   :-40.000   Min.   :-19.00   Min.   : 991   SW: 9387  
 1st Qu.: 29.00   1st Qu.:-10.000   1st Qu.:  2.00   1st Qu.:1008   NE: 4997  
 Median : 72.00   Median :  2.000   Median : 14.00   Median :1016   NW:14150  
 Mean   : 98.61   Mean   :  1.817   Mean   : 12.45   Mean   :1016   SE:15290  
 3rd Qu.:137.00   3rd Qu.: 15.000   3rd Qu.: 23.00   3rd Qu.:1025             
 Max.   :994.00   Max.   : 28.000   Max.   : 42.00   Max.   :1046             
 NA's   :2067                                                                 
      Iws                Is                Ir         
 Min.   :  0.45   Min.   : 0.00000   Min.   : 0.0000  
 1st Qu.:  1.79   1st Qu.: 0.00000   1st Qu.: 0.0000  
 Median :  5.37   Median : 0.00000   Median : 0.0000  
 Mean   : 23.89   Mean   : 0.05273   Mean   : 0.1949  
 3rd Qu.: 21.91   3rd Qu.: 0.00000   3rd Qu.: 0.0000  
 Max.   :585.60   Max.   :27.00000   Max.   :36.0000  
```



Normalizing the numeric attributes

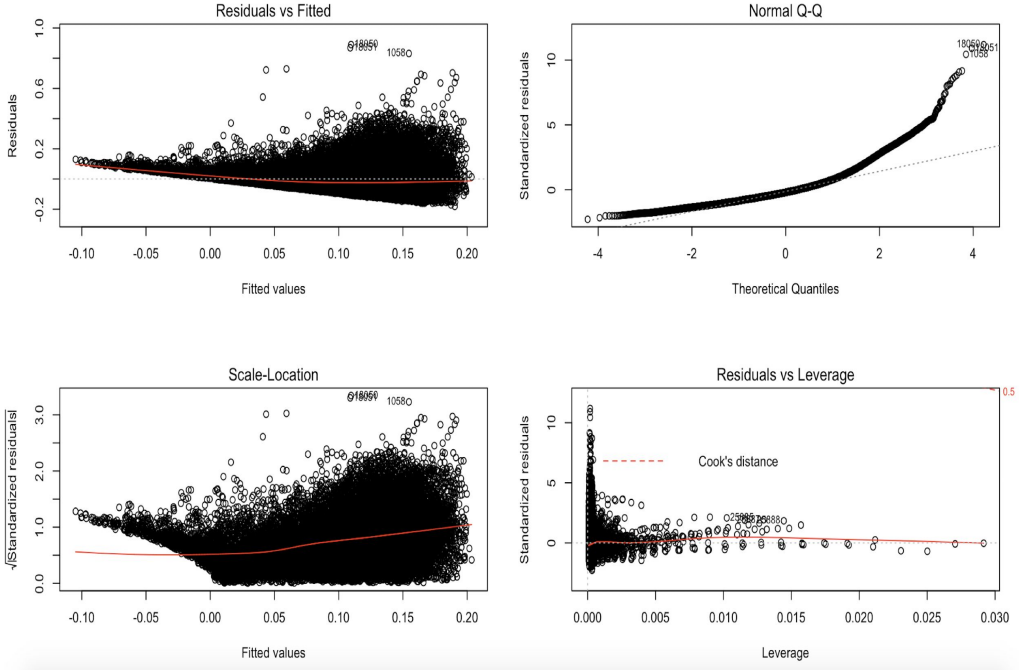
Omit the NA values

```
> summary(data)
      DEWP             TEMP             PRES              Iws          
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000  
 1st Qu.:0.4412   1st Qu.:0.3443   1st Qu.:0.3091   1st Qu.:0.002371  
 Median :0.6176   Median :0.5410   Median :0.4545   Median :0.008707  
 Mean   :0.6140   Mean   :0.5148   Mean   :0.4626   Mean   :0.041443  
 3rd Qu.:0.8088   3rd Qu.:0.6885   3rd Qu.:0.6182   3rd Qu.:0.037980  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000  
       Is                Ir              cwbd           pm2.5       
 Min.   :0.00000   Min.   :0.000000   cv: 8944   Min.   :  0.00  
 1st Qu.:0.00000   1st Qu.:0.000000   NE: 4756   1st Qu.: 29.00  
 Median :0.00000   Median :0.000000   NW:13484   Median : 72.00  
 Mean   :0.00205   Mean   :0.005413   SE:14573   Mean   : 98.61  
 3rd Qu.:0.00000   3rd Qu.:0.000000              3rd Qu.:137.00  
 Max.   :1.00000   Max.   :1.000000              Max.   :994.00
```



Multivariate Regression & Diagnostics

```
lm(formula = pm2.5 ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-180.23  -51.32  -15.80   31.36  885.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  185.595      3.848  48.227  < 2e-16 ***
DEWP         273.674      3.620  75.597  < 2e-16 ***
TEMP        -381.475      4.148 -91.969  < 2e-16 ***
PRES         -88.867      3.868 -22.972  < 2e-16 ***
Iws         -113.756      4.945 -23.003  < 2e-16 ***
Is           -89.117     13.576  -6.564 5.28e-11 ***
Ir          -224.288     10.019 -22.387  < 2e-16 ***
cwbdNE       -26.741      1.428 -18.722  < 2e-16 ***
cwbdNW       -30.108      1.173 -25.667  < 2e-16 ***
cwbdSE         3.882      1.092   3.555 0.000378 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.16 on 41747 degrees of freedom
Multiple R-squared:  0.2606,    Adjusted R-squared:  0.2604
F-statistic:  1635 on 9 and 41747 DF,  p-value: < 2.2e-16
```



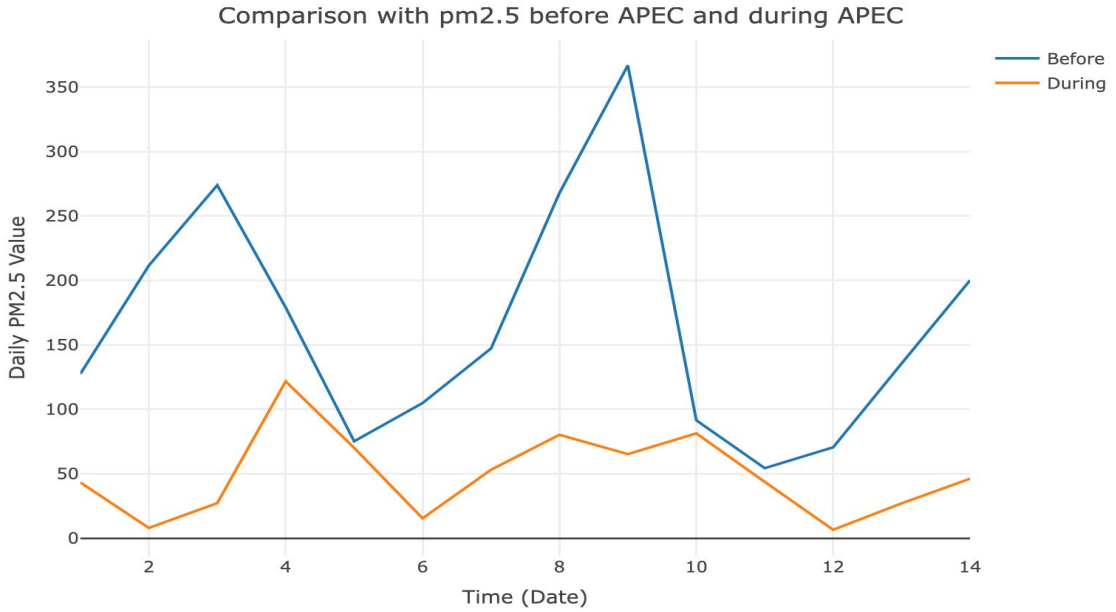
Results

- The distinct patterns of residuals vs fitted model indicates that linear assumptions is not good enough.
- The normal Q-Q plot examine whether the residuals are normally distributed, clearly the tail is far away from straight dashed line. So we have heavy tail.
- Scale-Location plot is used to check the homogeneity of the variance of the residuals. It’s good because we see a horizontal line with equally spread points.
- But we have 3 outliers whose standardized residuals are greater than 3 in absolute value.

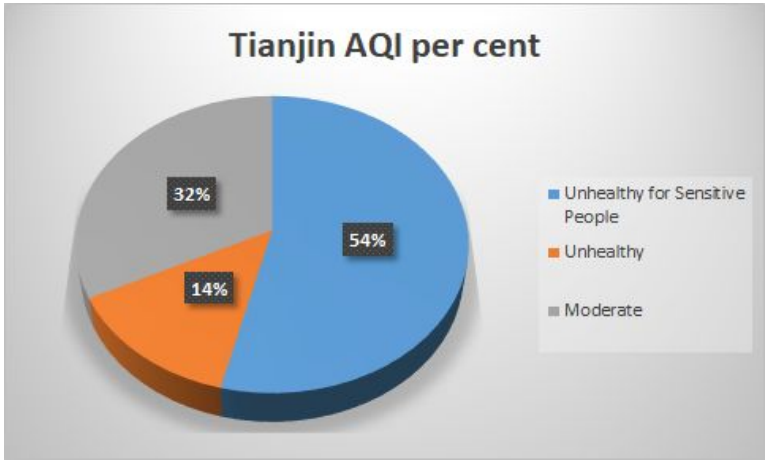
Suggestions to improve the air quality of Beijing

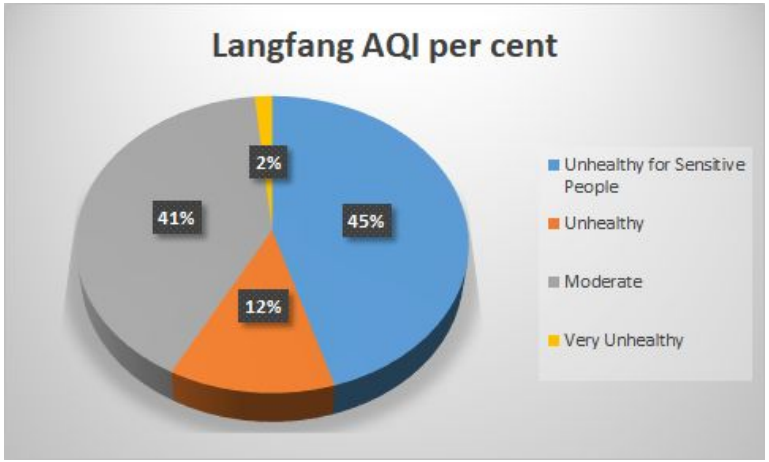
1. Take Special Administrative Measures

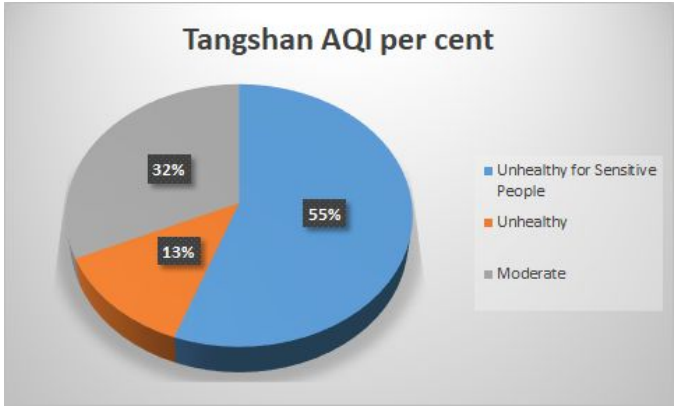
a. Traffic restrictions of based on the last digit of license plate numbers
b. License-plate lottery
c. Congestion fee
d. Closing high polluting factories around Beijing



2. Establish a Beijing-Tianjin-Hebei Air Quality Management District





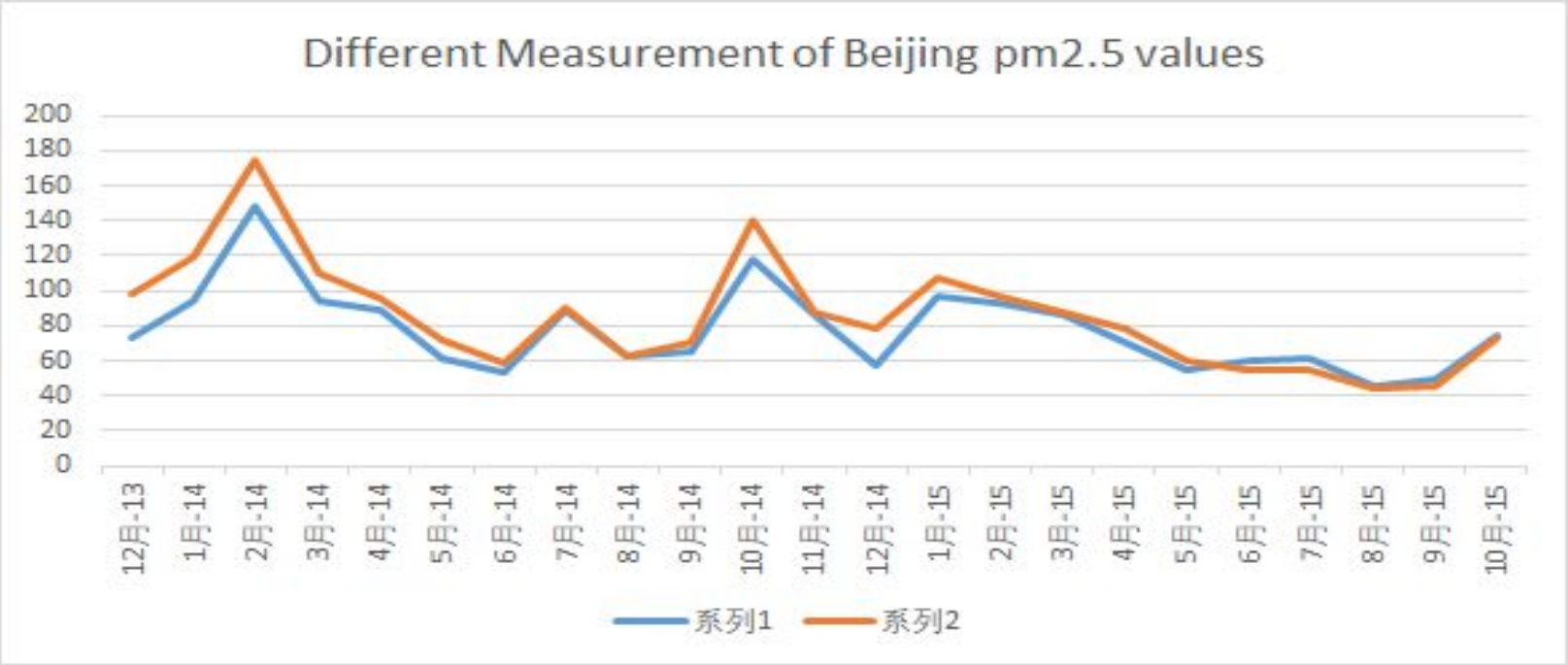


South Coast Air Quality Management District (SCAQMD)

Headquarters	Diamond Bar, CA United States
Key people	Wayne Nastri, Executive Officer
Website	http://www.aqmd.gov/

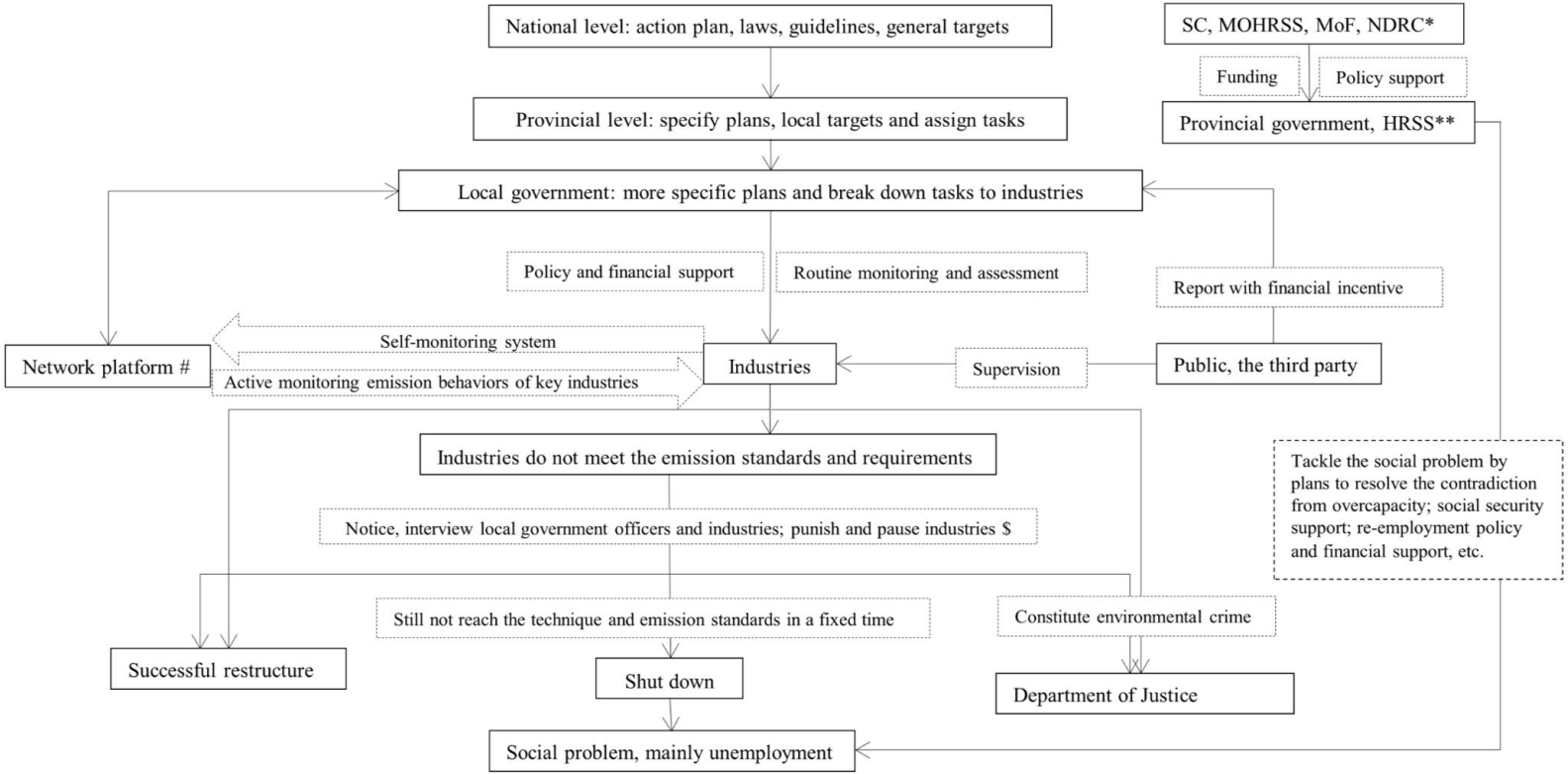
The South Coast Air Quality Management District, also using the acronym SCAQMD, formed in 1976, is the air pollution agency responsible for regulating stationary sources of air pollution in the South Coast Air Basin, in Southern California. The separate California Air Resources Board is responsible for regulating mobile sources (e.g. vehicles) in the air basin.

3. Establish NGO to collect the air pollution data



Conclusions and Next Steps for the Air Quality Crisis

Collaboration on Air Pollution Control



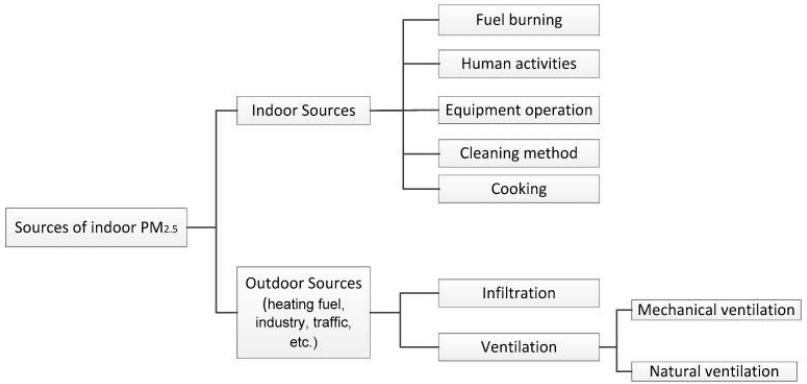
As noticed, Beijing Air Quality solutions are **complicated** and **take time**

INDOOR PM2.5

- Easier to address in the short term
- Development of New Material for Air Filters
- Anti-Haze Room Air Conditioners Available in the Market
- Note that outdoor pm2.5 can enter the room three ways: natural ventilation, mechanical ventilation, mechanical ventilation and infiltration

OUTDOOR PM2.5

- Heating, fuel, industry, traffic, etc.
- Rapid industrialization, high energy consumption, and large portion of coal (60-70%) in the structure of energy sources in China



Reference

- Taking Action on Air Pollution Control in the Beijing-Tianjin-Hebei (BTH) Region: Progress, Challenges and Opportunities
- World Economic Forum - “These are the world’s most polluted cities”, Tom Miles, Journalist at Reuters (03 May 2018)
- Quartz - “Six Years of Beijing air pollution summed up in one scary chart”, Lily Kuo
- US National Library of Medicine National Institutes of Health - “A Review of Recent Advances in Research on PM2.5 in China, Yaolin Lin, Jiale Zou, Wei Yang, Chun-Qing Li