Final_Project_Yifei_Zhang

https://github.com/yz470/Final-project-for-Environmental-Data-Analytics.git

Yifei Zhang

Abstract

This project aims at answering the following research questions: 1. Is there a trend in PM2.5 concentrations in Shanghai from 2012 to 2017? 2. What is the spatial pattern of PM2.5 concentration in Beijing, Chengdu, Guangzhou, Shanghai, and Shenyang in 2017? The U.S. Department of State Air Quality Dataset is used to answer the research questions. Results show that from 2012 to 2017, there existed a decreasing trend in PM2.5 concentrations in Shanghai. In addition, pronounced seasonal variation is observed in PM2.5 concentrations in Shanghai, with the highest concentrations typically observed in the winter and the lowest concentrations generally found in the summer. Among the 5 cities, Beijing and Chengdu showed highest average PM2.5 concentration in 2017, with the AQI level reaching 158.

Contents

List of Tables

List of Figures

1 Research Question and Rationale

Air pollution is a severe problem in China. It is often the smoggy air that occurs to many people when it comes to Beijing, the capital of China. Among the air pollutants, of particular concern is PM2.5 (particles with an aerodynamic diameter less than 2.5 um). Exposure to high concentrations of PM2.5 results in risks to the cardiovascular system, cerebrovascular system and an increase in the probability of cancer and premature death. In recent years, China's government has taken serious action against air pollution, but there is still a long way to go.

Given the increasing public concern in air quality and the efforts of Chinese government to fight air pollution, I am interested in the actual outcomes: whether or not there is a decreasing trend in air quality over the years. In particular, I want to examine the PM2.5 trend in Shanghai because it is the largest city in China by population, and it is also where I come from. I will use the The U.S. Department of State air quality files, which contain the hourly PM2.5 concentrations in 5 cities in China from 2011 to 2017. With this dataset, my second research question is to look at the spatial pattern of PM2.5 concentrations in these 5 cities.

Goals:

- Use Mann-Kendall test to analyze trends in PM2.5 concentrations in Shanghai from 2012 to 2017
- Run a Pettitt's Test to check if there are changing points
- Check for seasonality in PM2.5 concentrations.
- Use Seasonal Mann-Kendall test if there is seasonality
- Look at which city has the highest average PM2.5 concentration.

2 Dataset Information

The U.S. Department of State air quality files contain hourly PM2.5 or data in concentration units from each post, as reported on the www.stateair.net website. Files include hourly data with the following file name structure: Site_Year_DurationParameter.csv. Filename examples: Beijing_2013_HourlyPM2.5.csv

All files contain the following column headers: Site, Parameter, Date (LST), Year, Month, Day, Hour, Value, Unit, Duration, QC Name. Definitions and examples of column headers can be found in the table below.

The air quality data are measured at the U.S. Embassy in Beijing and at the Consulates in Chengdu, Guangzhou, Shanghai, and Shenyang.

Table 1: Data Structure and definition

Term	Definitions	Examples
Site	City or post where the measurements were taken.	Beijing, Shenyang
Parameter	The air quality pollutant measured.	PM2.5, O3
Date	The date and hour of the measurement in local standard time (e.g., BJT – Beijing Time). The date-time format follows YYYY-MM-DD HH:mm, where 00:00 is midnight, 14:00 is 2:00 p.m., etc.	2013-05-01 00:00
Year	4 digit year that corresponds to YYYY in Date	2013
Month	1 or 2 digit month (1 to 12) that corresponds to MM in Date	5, 12
Day	1 or 2 digit day (1 to 31) that corresponds to DD in Date	1, 31
Hour	1 or 2 digit hour (0 to 23) that corresponds to HH in Date	0, 18
Value	The measurement in concentration. Missing values are listed as -999.	45, 450, -999
Unit	(ug/m3) for PM2.5	ug/m3
Duration	1-hour (1 Hr) for PM2.5	1 Hr
QC Name	The quality control status of the data; either valid or missing (unavailable). Invalid data are not included in these files.	Valid, Missing

Table 2: Embassy and Consulate geographic coordinates.

Site location	Latitude and Longitude Degrees
Beijing	39.95, 116.47
Chengdu	30.63, 104.07
Guangzhou	23.12, 113.32
Shanghai	31.21, 121.44
Shenyang	41.78, 123.42

3 Exploratory Data Analysis and Wrangling

```r
#Look at the general structure of the data, number of observations, maximum and minimu
colnames(shanghai_2017)
```

```
##  [1] "Site"         "Parameter"   "Date..LST." "Year"         "Month"
##  [6] "Day"          "Hour"        "Value"      "Unit"         "Duration"
## [11] "QC.Name"
```

```r
dim(shanghai_2017)
```

```
## [1] 4344   11
```

```r
head(shanghai_2017)
```

```
##       Site Parameter    Date..LST. Year Month Day Hour Value  Unit
## 1 Shanghai     PM2.5 1/1/2017 0:00 2017     1   1    0    42 µg/m³
## 2 Shanghai     PM2.5 1/1/2017 1:00 2017     1   1    1    46 µg/m³
## 3 Shanghai     PM2.5 1/1/2017 2:00 2017     1   1    2    56 µg/m³
## 4 Shanghai     PM2.5 1/1/2017 3:00 2017     1   1    3    49 µg/m³
## 5 Shanghai     PM2.5 1/1/2017 4:00 2017     1   1    4    51 µg/m³
## 6 Shanghai     PM2.5 1/1/2017 5:00 2017     1   1    5    49 µg/m³
##   Duration QC.Name
## 1     1 Hr   Valid
## 2     1 Hr   Valid
## 3     1 Hr   Valid
## 4     1 Hr   Valid
## 5     1 Hr   Valid
## 6     1 Hr   Valid
```

```r
str(shanghai_2017)
```

```
## 'data.frame':    4344 obs. of  11 variables:
##  $ Site      : Factor w/ 1 level "Shanghai": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Parameter : Factor w/ 1 level "PM2.5": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date..LST.: Factor w/ 4343 levels "1/1/2017 0:00",..: 1 2 13 18 19 20 21 22 23 24
##  $ Year      : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ Month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Hour      : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Value     : int  42 46 56 49 51 49 43 47 52 45 ...
##  $ Unit      : Factor w/ 1 level "µg/m³": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Duration  : Factor w/ 1 level "1 Hr": 1 1 1 1 1 1 1 1 1 1 ...
##  $ QC.Name   : Factor w/ 2 levels "Missing","Valid": 2 2 2 2 2 2 2 2 2 2 ...
```

```r
summary(shanghai_2017)
```

```
##       Site       Parameter              Date..LST.         Year
##  Shanghai:4344   PM2.5:4344   3/12/2017 3:00:   2   Min.   :2017
##                               1/1/2017 0:00 :   1   1st Qu.:2017
##                               1/1/2017 1:00 :   1   Median :2017
##                               1/1/2017 10:00:   1   Mean   :2017
##                               1/1/2017 11:00:   1   3rd Qu.:2017
##                               1/1/2017 12:00:   1   Max.   :2017
##                               (Other)       :4337
##      Month            Day            Hour            Value
##  Min.   :1.000   Min.   : 1.0   Min.   : 0.00   Min.   :-999.00
##  1st Qu.:2.000   1st Qu.: 8.0   1st Qu.: 5.75   1st Qu.:  24.00
##  Median :4.000   Median :16.0   Median :11.50   Median :  39.00
##  Mean   :3.508   Mean   :15.6   Mean   :11.50   Mean   :  27.87
##  3rd Qu.:5.000   3rd Qu.:23.0   3rd Qu.:17.25   3rd Qu.:  59.00
##  Max.   :6.000   Max.   :31.0   Max.   :23.00   Max.   : 188.00
##
##     Unit       Duration       QC.Name
##  µg/m³:4344   1 Hr:4344   Missing:  73
##                           Valid  :4271
##
##
##
##
##
```

```r
summary(shanghai_2012)
```

```
##       Site       Parameter                Date..LST.         Year
##  Shanghai:8784   PM2.5:8784   2012-03-11 03:00:   2   Min.   :2012
##                               2012-01-01 00:00:   1   1st Qu.:2012
##                               2012-01-01 01:00:   1   Median :2012
##                               2012-01-01 02:00:   1   Mean   :2012
##                               2012-01-01 03:00:   1   3rd Qu.:2012
##                               2012-01-01 04:00:   1   Max.   :2012
##                               (Other)         :8777
##      Month             Day             Hour            Value
##  Min.   : 1.000   Min.   : 1.00   Min.   : 0.00   Min.   :-999.00
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 5.75   1st Qu.:  23.00
##  Median : 7.000   Median :16.00   Median :11.50   Median :  40.00
##  Mean   : 6.514   Mean   :15.76   Mean   :11.50   Mean   :  17.21
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:17.25   3rd Qu.:  65.00
##  Max.   :12.000   Max.   :31.00   Max.   :23.00   Max.   : 650.00
##
##     Unit       Duration       QC.Name
##  µg/m³:8784   1 Hr:8784   Missing:   5
```

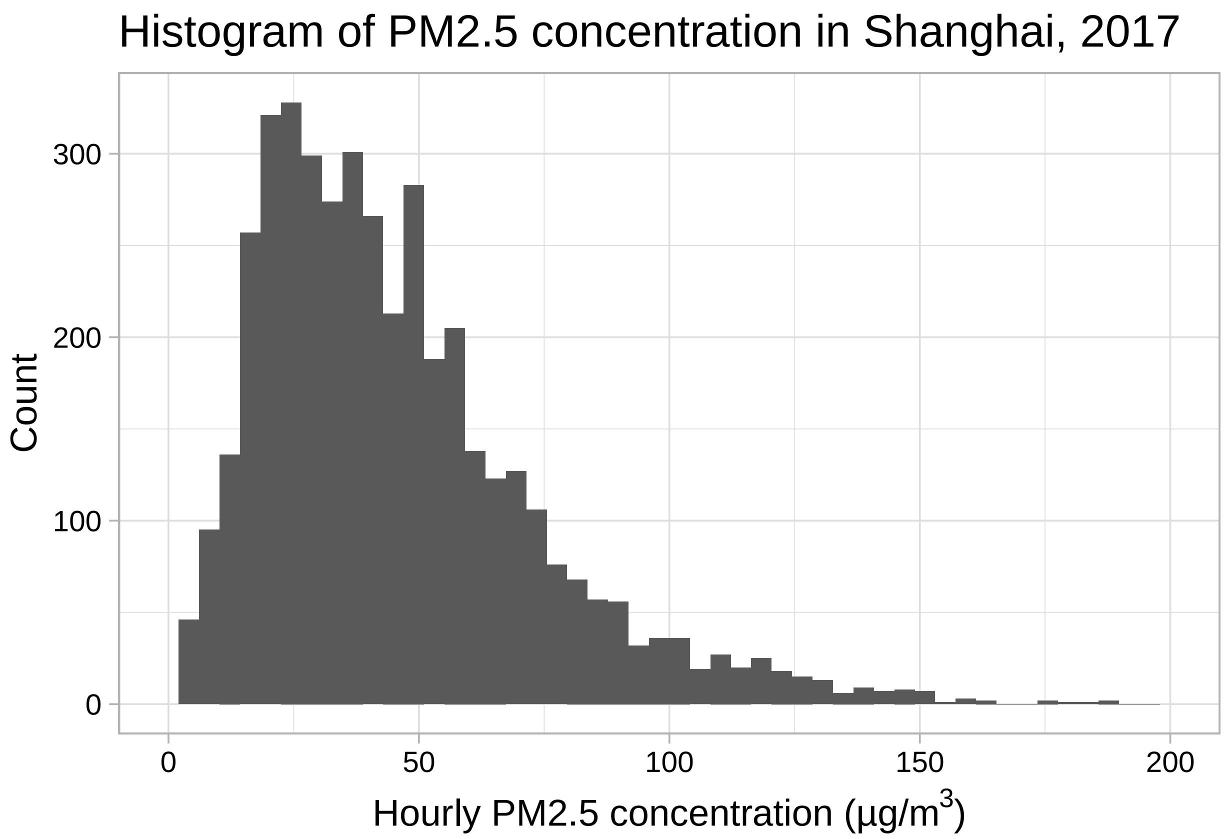

Figure 1: Histogram of PM2.5 concentration in Shanghai, 2017

```
##                            Valid  :8779
##
##
##
##
##
```

Figure 1 shows how PM2.5 concentrations spread over the range of values

Figure 3 Compares the mean PM2.5 concentrations of 2012 and 2017

```
#combine Shanghai data into one dataframe
Shanghai12to17 <- do.call("rbind", list(shanghai_2012, shanghai_2013, shanghai_2014, sha

#filter out missing values and calculate daily average PM2.5 concentration
Shanghai12to17_Processed <- Shanghai12to17 %>%
  filter(QC.Name == "Valid" & Value >=0)%>%
  group_by(Year,Month, Day)%>%
  summarise(Daily_Mean_PM25 = mean(Value))
```

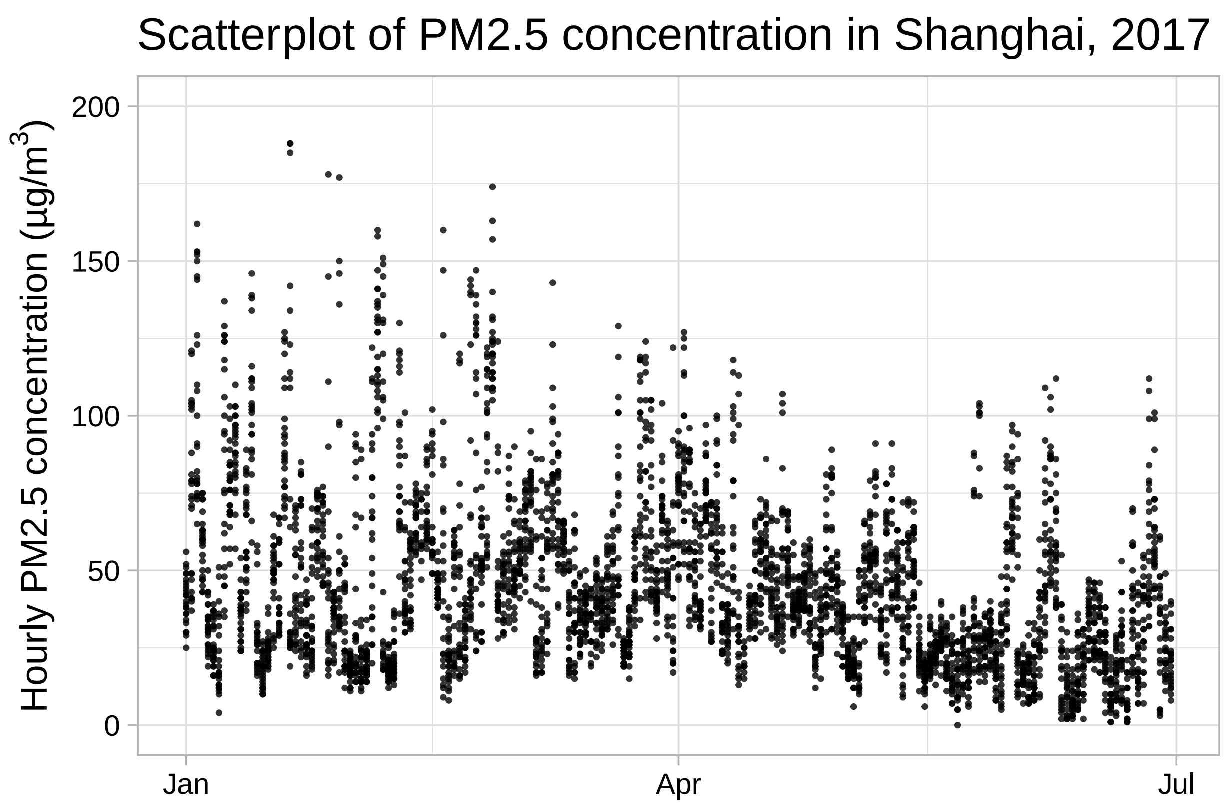

Figure 2: Scatterplot of PM2.5 concentration in Shanghai, 2017

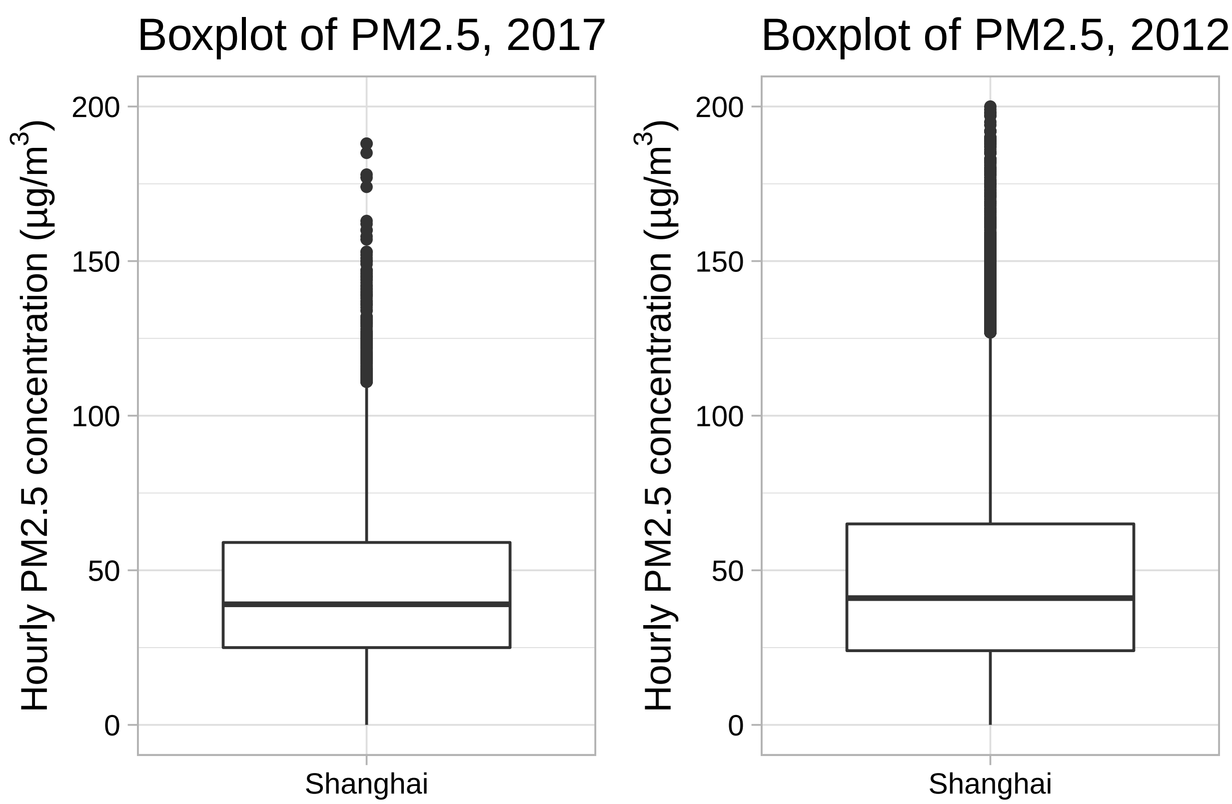

Figure 3: Boxplot of PM2.5 concentration in Shanghai, 2017 and 2012

```r
#Create Date
  Shanghai12to17_Processed$Date <- as.Date(with(Shanghai12to17_Processed, paste(Year, Mo

#Rearrange columns
  Shanghai12to17_Processed <- Shanghai12to17_Processed[,c(5,1,2,3,4)]
```

```r
#Assign coordinates to each site
Beijing_2017$Latitude <- 39.95
  Beijing_2017$Longitude <- 116.47

Chengdu_2017$Latitude <- 30.63
  Chengdu_2017$Longitude <- 104.07

Guangzhou_2017$Latitude <- 23.12
  Guangzhou_2017$Longitude <- 113.32

Shenyang_2017$Latitude <- 41.78
  Shenyang_2017$Longitude <-123.42

shanghai_2017$Latitude <- 31.21
  shanghai_2017$Longitude <- 121.44

#Combine data for all sites into one dataframe
All_Sites_2017 <- do.call("rbind", list(shanghai_2017, Shenyang_2017, Guangzhou_2017, Ch

#Filter out missing value and calculate daily average PM2.5 concentrations
All_Sites_2017_Processed <- All_Sites_2017 %>%
  filter(QC.Name == "Valid" & Value >=0)%>%
  group_by(Site, Year, Month, Day, Latitude, Longitude)%>%
  summarise(Daily_Mean_PM25 = mean(Value))

All_Sites_2017_avg <- All_Sites_2017_Processed %>%
  group_by(Site, Latitude, Longitude) %>%
  summarise(meanPM25 = mean(Daily_Mean_PM25),
            maxPM25 = max(Daily_Mean_PM25))
```

```r
#Create a simplified function to convert PM2.5 concentration to AQI
Concentration_to_AQI <- function(C){
  if (C <= 55.4){
    as.integer((C-35.5)*(150-101)/(55.4-35.5)+101)
  } else{
    if (C <= 150.4){
      as.integer((C-55.5)*(200-151)/(150.4-55.5)+151)
    }
  }
```

```
}

All_Sites_2017_avg <- mutate(All_Sites_2017_avg, AQI = Concentration_to_AQI(meanPM25))
```

4 Analysis

```r
#run a Mann-Kendall test on the PM2.5 concentration in Shanghai from 2012 to 2017
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25)
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25
## z = -4.6494, n = 1995, p-value = 3.329e-06
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.381510e+05  8.829007e+08 -6.947051e-02
```

```r
#Run a Pettitt's Test to check if there are changing points
pettitt.test(Shanghai12to17_Processed$Daily_Mean_PM25)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25
## U* = 145860, p-value = 2.103e-07
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                            1597
```

```r
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1:1597])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1:1597]
## z = -0.1021, n = 1597, p-value = 0.9187
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## -2.17400e+03  4.52980e+08 -1.70621e-03
```

```r
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1598:1995])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1598:1995]
```

```
## z = 3.7256, n = 398, p-value = 0.0001948
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 9.880000e+03 7.031225e+06 1.250910e-01
```

```r
# Is there a second change point?
pettitt.test(Shanghai12to17_Processed$Daily_Mean_PM25[1598:1995])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1598:1995]
## U* = 19508, p-value = 4.084e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             158
```

```r
# Run another Mann-Kendall for the second change point
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1598:1755])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1598:1755]
## z = -4.5497, n = 158, p-value = 5.373e-06
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -3.027000e+03  4.423650e+05 -2.441326e-01
```

```r
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1756:1995])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1756:1995]
## z = -5.3089, n = 240, p-value = 1.103e-07
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -6.601000e+03  1.545518e+06 -2.302206e-01
```

```r
pettitt.test(Shanghai12to17_Processed$Daily_Mean_PM25[1598:1755])
```

```
##
##  Pettitt's test for single change-point detection
```

```
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1598:1755]
## U* = 3356, p-value = 8.076e-08
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               65
```

```
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1598:1662])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1598:1662]
## z = 0.40196, n = 65, p-value = 0.6877
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 7.200000e+01 3.120000e+04 3.461538e-02
```

```
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1663:1755])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1663:1755]
## z = 0.84963, n = 93, p-value = 0.3955
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.570000e+02 9.078567e+04 6.009588e-02
```

```
pettitt.test(Shanghai12to17_Processed$Daily_Mean_PM25[1756:1995])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1756:1995]
## U* = 6169, p-value = 1.436e-07
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              170
```

```
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1756:1925])
```

```
##
##  Mann-Kendall trend test
```

```
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1756:1925]
## z = 0.014824, n = 170, p-value = 0.9882
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.200000e+01 5.506513e+05 8.355673e-04
```

```
mk.test(Shanghai12to17_Processed$Daily_Mean_PM25[1926:1995])
```

```
##
##  Mann-Kendall trend test
##
## data:  Shanghai12to17_Processed$Daily_Mean_PM25[1926:1995]
## z = -2.2459, n = 70, p-value = 0.02471
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS            tau
##  -444.0000000 38905.3333333    -0.1839652
```

According to the first Mann-Kendall Test, p-value < 0.0001, S = -1.38e+05 so there is a decreasing trend in PM2.5 concentrations in Shanghai from 2012 to 2017

```
Shanghai12to17_Monthly <- Shanghai12to17_Processed %>%
  group_by(Year,Month)%>%
  summarise(Monthly_Mean_PM25 = mean(Daily_Mean_PM25))

Shanghai12to17_Monthly$Date <- as.Date(with(Shanghai12to17_Monthly, paste(Year, Month,"(

# Create a time series object
Shanghai12to17_timeseries <- ts(Shanghai12to17_Monthly$Monthly_Mean_PM25,
                                start = c(2012, 1) ,frequency = 12)
Shanghai12to17_timeseries
```

```
##            Jan       Feb       Mar       Apr       May       Jun       Jul
## 2012  64.39169  53.10445  67.14090  54.95019  55.70430  40.16078  26.55297
## 2013 101.46242  64.25839  64.96774  61.61860  53.23069  44.33886  32.14922
## 2014  72.61396  52.32508  54.39317  51.65912  61.19360  42.12621  35.17865
## 2015  84.64382  66.46796  50.98786  52.40304  40.98860  36.55009  35.35663
## 2016  72.58816  57.58583  55.78133  56.02394  48.52583  35.32621  34.72419
## 2017  52.58442  56.67838  51.04167  49.21948  32.92391  30.73789
##            Aug       Sep       Oct       Nov       Dec
## 2012  16.12474  44.06365  55.51799  67.04937  64.37512
## 2013  26.45089  27.97110  36.20997  79.18295 122.15358
## 2014  29.41896  29.09568  38.79696  53.27199  74.75655
## 2015  35.41667  28.07370  38.43414  55.90320  82.21746
## 2016  20.21237  28.34902  20.96657  50.00350  64.36563
```

```
## 2017
```

```r
# Run a Seasonal Mann-Kendall test
Shanghai.smktest <- smk.test(Shanghai12to17_timeseries)
Shanghai.smktest
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  Shanghai12to17_timeseries
## z = -2.6169, p-value = 0.008873
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##    S varS
##  -44  270
```

```r
summary(Shanghai.smktest)
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: Shanghai12to17_timeseries
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                        S varS    tau      z Pr(>|z|)
## Season 1:   S = 0   -5 28.3 -0.333 -0.751 0.452370
## Season 2:   S = 0    1 28.3  0.067  0.000 1.000000
## Season 3:   S = 0   -9 28.3 -0.600 -1.503 0.132855
## Season 4:   S = 0   -5 28.3 -0.333 -0.751 0.452370
## Season 5:   S = 0   -9 28.3 -0.600 -1.503 0.132855
## Season 6:   S = 0  -11 28.3 -0.733 -1.879 0.060289 .
## Season 7:   S = 0    6 16.7  0.600  1.225 0.220671
## Season 8:   S = 0    4 16.7  0.400  0.735 0.462433
## Season 9:   S = 0   -2 16.7 -0.200 -0.245 0.806496
## Season 10:   S = 0  -6 16.7 -0.600 -1.225 0.220671
## Season 11:   S = 0  -6 16.7 -0.600 -1.225 0.220671
## Season 12:   S = 0  -2 16.7 -0.200 -0.245 0.806496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#See if there is a change point.
pettitt.test(Shanghai12to17_timeseries)
```

```
##
```

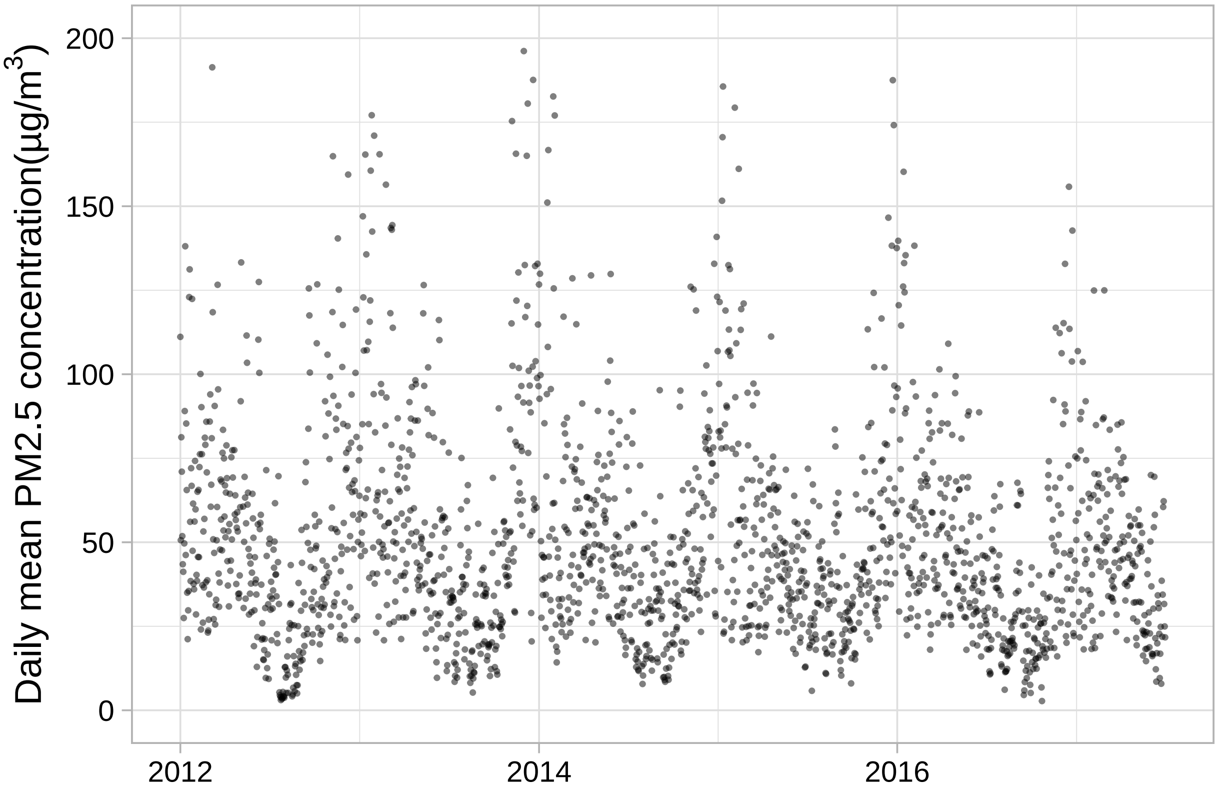

Figure 4: Initial visualization of Shanghai PM2.5 concentration from 2012 to 2017

```
##  Pettitt's test for single change-point detection
##
## data:  Shanghai12to17_timeseries
## U* = 296, p-value = 0.3302
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               52
```

The changing points in figure 5 seem to be due to the seasonal variations, so I decide to look at the monthly variation of PM2.5 concentration to see if there exists seasonality

Figure 6 shows that the PM2.5 concentraions do vary a lot over months. Generally the concentrations peak in Winter and reach lowest in Summer. Therefore, I decide to run a Seasonal Mann-Kendall test to reduce to effect of seasonal trends that obscure the overall direction of the trend

Figure 7 shows the average PM2.5 concentration in 5 cities in China in 2017. The larger the circle, the higher the concentration. In the HTML format, the exact AQI value will pop up if the circle is clicked, but this function is not available in the PDF format.

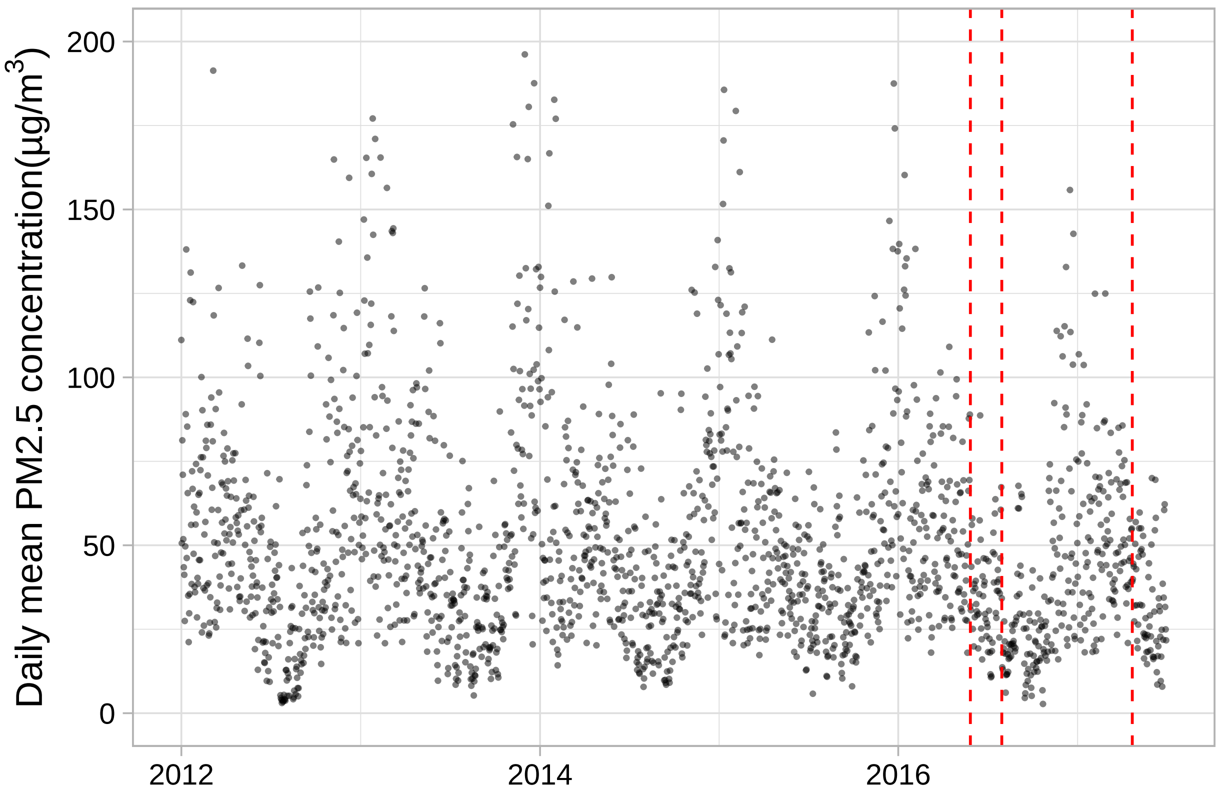

Figure 5: Shanghai PM2.5 concentration from 2012 to 2017 with changing points

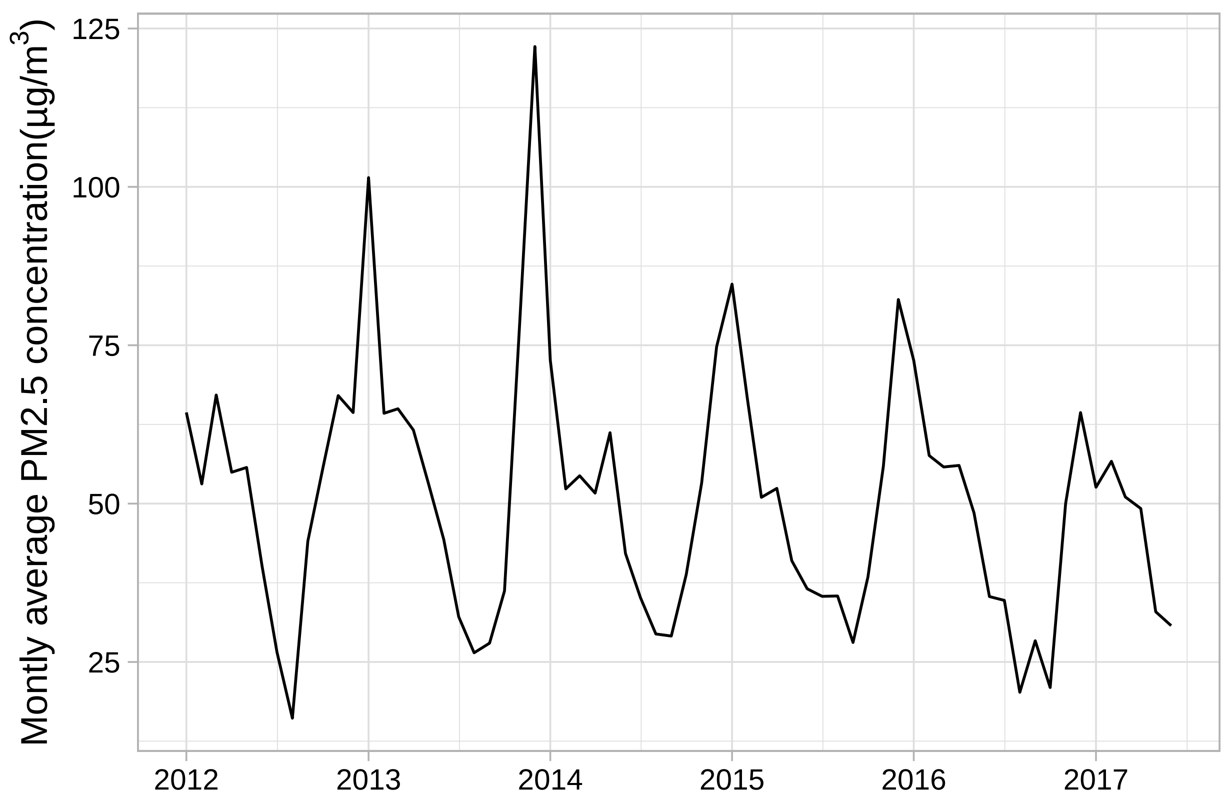

Figure 6: Shanghai Monthly average PM2.5 concentration from 2012 to 2017

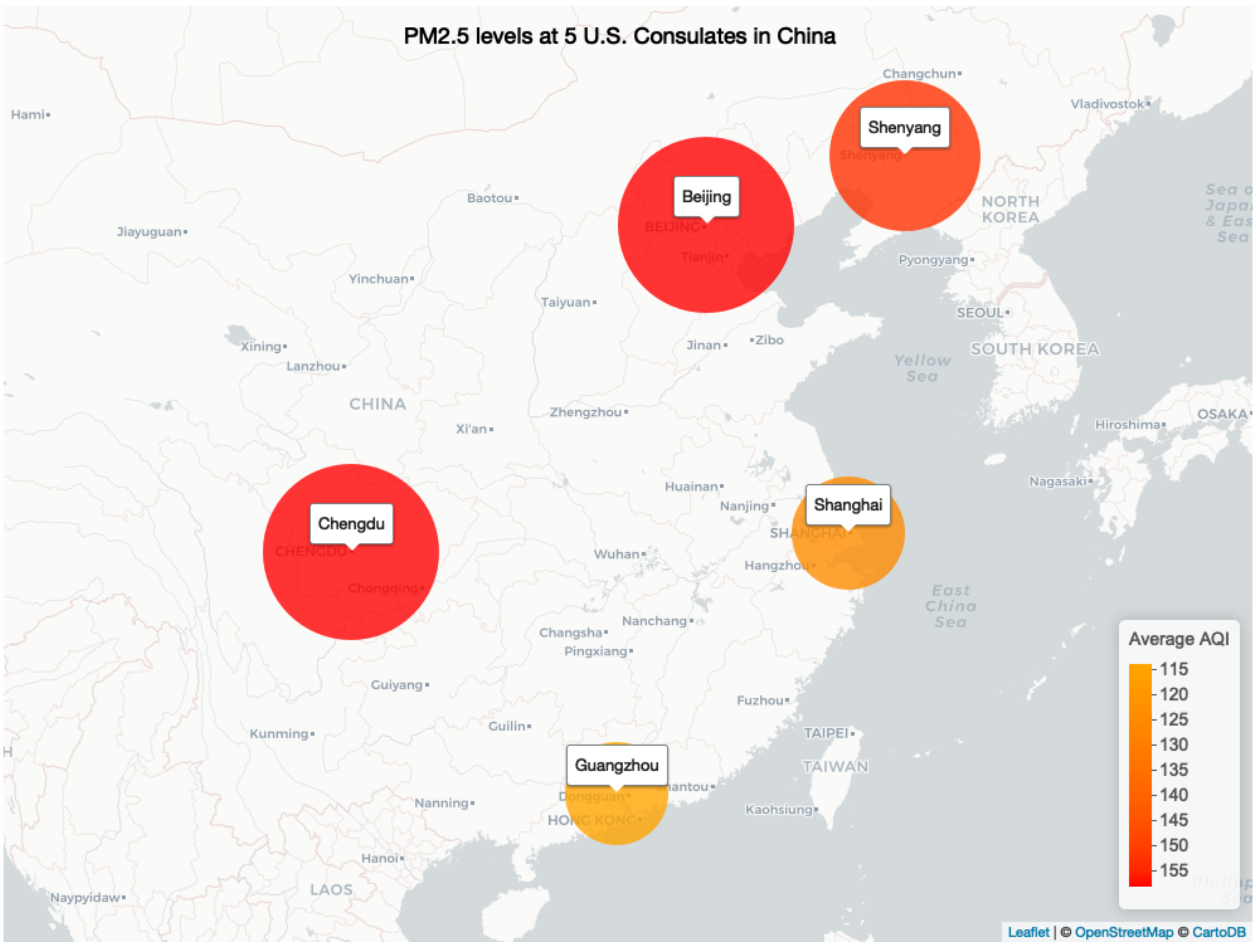

Figure 7: PM2.5 levels at 5 U.S. Consulates in China

5 Summary and Conclusions

Pronounced seasonal variation is observed in PM2.5 concentrations in Shanghai, with the highest concentrations typically observed in the winter and the lowest concentrations generally found in the summer. From 2012 to 2017, there existed a decreasing trend in PM2.5 concentrations in Shanghai(Seasonal Mann-Kendall test, p-value = 0.008873, S = -44). China's government has made some progress to improve air quality. However, the 2017 average AQI in Shanghai was 125, which falls into the "unhealthy for sensitive groups" range.

As for other 4 cities, Beijing and Chengdu showed highest PM2.5 concentration in 2017, with the AQI level reaching 158. Guangzhou showed lowest AQI in 2017, which is 114; the 2017 average AQI in Shenyang was 153, also within the unhealthy range. Transportation and Coal combustion are two major sources of PM2.5 in China. Transportation contributes a lot to the PM2.5 concentration in Beijing and Chengdu. These two cities rank first and second in terms of car ownership in China.