Was $PM_{2.5}$ concentration in Beijing decreasing?

Xiaozhu Zhang

Report submitted for the final project of
Introduction to Data Science

UCLA
June 2019

Contents

1 Introduction

1.1 Project description

Before the Beijing 2008 Summer Olympic Games, the severe air pollution in Beijing had already caught the whole world's attention. Among all those gauges of air pollution, $\mathrm{PM}_{2.5}$ concentration is no doubt the most famous and effective. According to U.S. Environmental Protection Agency (EPA), the air quality index (AQI) could be regarded as "Good" if AQI is less than 50, whereas the average AQI in Beijing from 2010 to 2014 was almost 100. These days, more people in China have been aware of the importance of clean environment, and Chinese government had made many policies in recent years to improve the air quality.

However, regardless of all those efforts made by citizens and the government, we would like to ask: what are the most influential factors contributing to $\mathrm{PM}_{2.5}$ concentration? How to predict it? And, most importantly, whether the air quality was getting better or not?

1.2 The data set

The dataset could be downloaded from UCI Repository with the link `https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data`. There are 13 variables in the original dataset:

No	Row number;
$year$	Year of data in this row;
$month$	Month of data in this row;
day	Day of data in this row;
$hour$	Hour of data in this row;
$pm2.5$	$\mathrm{PM}_{2.5}$ concentration (ug/m^3);
$DEWP$	Dew Point;
$TEMP$	Temperature;
$PRES$	Pressure (hPa);
$cbwd$	Combined wind direction;
Iws	Cumulated wind speed (m/s);
Is	Cumulated hours of snow;
Ir	Cumulated hours of rain.

Each variable is described by time series data from the year 2010 to 2014, in the unit of hour; therefore, there are 43,824 observations in total. There are no missing values except for the response variable $pm2.5$, 4.72% of whom are lost.

2 Preparation

2.1 Dealing with missing values

As mentioned above, there are missing values in the dataset. Since the exact variable with missing values is the response variable, here we delete all those rows with missing values directly. In fact, we could impute those missing values with the prediction algorithm we will use in the following section, and therefore we are able to answer the ultimate question that whether the air quality in Beijing was getting better or not.

2.2 Basic understanding of the dataset

Now, we would like to explore the dataset in order to reach a basic understanding of the distribution of continuous variables, and also the levels of categorical variables. We use "summary" to display the properties of the dataset.

```
> summary(dat)
```

```
   year          month             day              hour            pm2.5
 2010:8760   1      : 3720   1      : 1440   0      : 1826   Min.   :  0.00
 2011:8760   3      : 3720   2      : 1440   1      : 1826   1st Qu.: 29.00
 2012:8784   5      : 3720   3      : 1440   2      : 1826   Median : 72.00
 2013:8760   7      : 3720   4      : 1440   3      : 1826   Mean   : 98.61
 2014:8760   8      : 3720   5      : 1440   4      : 1826   3rd Qu.:137.00
             10     : 3720   6      : 1440   5      : 1826   Max.   :994.00
             (Other):21504   (Other):35184   (Other):32868   NA's   :2067
      DEWP               TEMP              PRES       cbwd          Iws
 Min.   :-40.000   Min.   :-19.00   Min.   : 991   cv: 9387   Min.   :  0.45
 1st Qu.:-10.000   1st Qu.:  2.00   1st Qu.:1008   NE: 4997   1st Qu.:  1.79
 Median :  2.000   Median : 14.00   Median :1016   NW:14150   Median :  5.37
 Mean   :  1.817   Mean   : 12.45   Mean   :1016   SE:15290   Mean   : 23.89
 3rd Qu.: 15.000   3rd Qu.: 23.00   3rd Qu.:1025              3rd Qu.: 21.91
 Max.   : 28.000   Max.   : 42.00   Max.   :1046              Max.   :585.60

       Is                Ir
 Min.   : 0.00000   Min.   : 0.0000
 1st Qu.: 0.00000   1st Qu.: 0.0000
 Median : 0.00000   Median : 0.0000
 Mean   : 0.05273   Mean   : 0.1949
 3rd Qu.: 0.00000   3rd Qu.: 0.0000
 Max.   :27.00000   Max.   :36.0000
```

In the original dataset, the only categorical variable about meteorological factors is *cbwd*, with 4 levels representing 4 different wind directions, which are CV(calm and variable), NE(northeast), NW(northwest) and SE(southeast). According to the table, the direction of NW and SE are more common than that of CV and NE.

Since all other variables are continuous, exploratory graphs could be shown to help us get a more vivid comprehension of the distributions. From Figure 1(a) to 1(g), we could tell that the distributions of variables $PRES$, $TEMP$ and $DEWP$ are more like the so-called "bell shape", whereas other variables have extremely fat tails. Therefore, transformation is needed to improve the quality of dataset.

We still have 4 more variables to tackle, all of whom are about the specific time of each observation. In fact, due to variation of meteorological factors, $\mathrm{PM}_{2.5}$ concentration may vary significantly in different seasons; $\mathrm{PM}_{2.5}$ concentration may also increase during rush hours every day. For the sake of simplicity and efficiency, we will transform the variables *month* and *hour* by merging some of their levels.

2.3 Data transformation

In this section, we will conduct the transformation of variables mentioned above in detail.

2.3.1 Logarithmic transformation

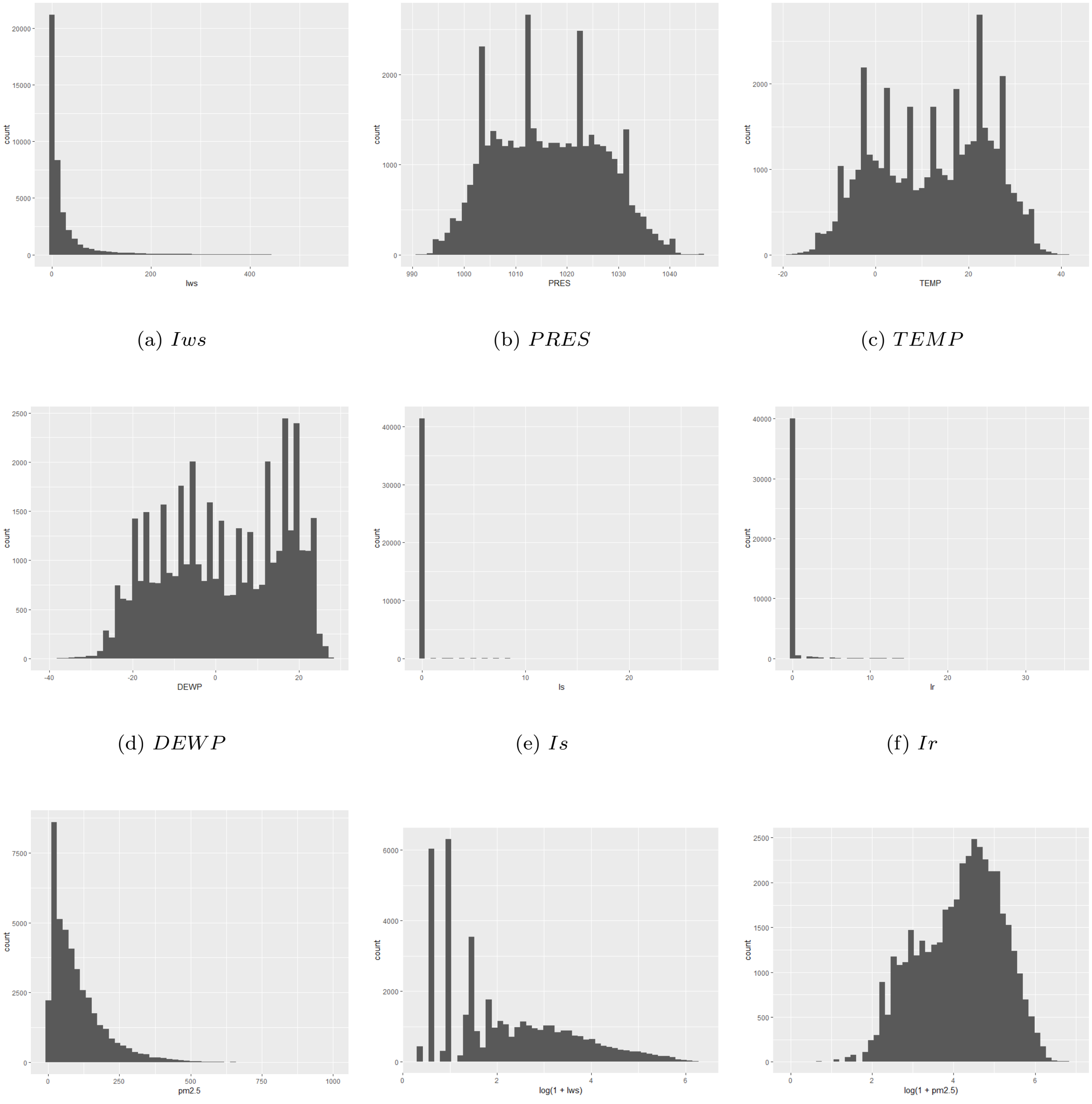
As mentioned above, the variable Iws and $pm2.5$ have extremely fat tails. Therefore, it could be a good way to take advantage of logarithmic transformation to generate new variables almost normally distributed:

$$Y = \log(1 + X)$$

where X is the original variable and Y is the new variable. After transformation, the distributions of $\log(1 + Iws)$ and $\log(1 + pm2.5)$ could be described as Figure 1(h) and 1(i), where the variables are roughly normally distributed. Therefore, in further analysis, we will use $\log(1 + Iws)$ and $\log(1 + pm2.5)$, instead of using variables Iws and $pm2.5$ directly.

2.3.2 Merging levels of categorical variables

There are 12 levels in the *month* variable and 24 levels in the *hour* variable, which are quite cumbersome if we input them into machine learning algorithms. What's more, the concentration of $\mathrm{PM}_{2.5}$ may not change significantly in a single month or a single hour. Only if we merge those finely-divided levels into some more



(a) Iws

(b) $PRES$

(c) $TEMP$

(d) $DEWP$

(e) Is

(f) Ir

(g) $pm2.5$

(h) $\log(1 + Iws)$

(i) $\log(1 + pm2.5)$

Figure 1: Histograms of continuous variables

rough levels are we more likely to observe a certain pattern between those predictors and the response variable.

The merging process is described as follows:

Table 1: Merging levels of categorical variables

Month	3, 4, 5	6, 7, 8	9, 10, 11	12, 1, 2
Season	Spring	Summer	Fall	Winter

Hour	1-6	7-12	13-18	19-24(0)
Time	Night	Morning	Afternoon	Evening

from which we derived 2 new categorical variables, *season* (4 levels) and *time* (4 levels). They are quite useful in the following analysis.

2.3.3 Cluster based on five-number-summary

In this part, we will deal with the variables Ir and Is. Notice that Ir denotes cumulative hours of rain and Is denotes cumulative hours of snow. Since Beijing is in a relatively arid geographical area, most values of the two variables are 0. What's more, we also notice that,

```
> library(sqldf)
> sqldf('SELECT * FROM dat WHERE [Is] != 0 and [Ir] !=0')     # no such row exists

 [1] year   month  day    hour   pm2.5  DEWP   TEMP   PRES   cbwd   Iws
[11] Is     Ir     season time
<0 rows> (or 0-length row.names)
```

indicating that there is no overlapping of nonzero values between the two variables. For the sake of convenience in further analysis, we simply create a new variable:

$$PRECIP = Is + Ir$$

which represents precipitation including both rain and snow.

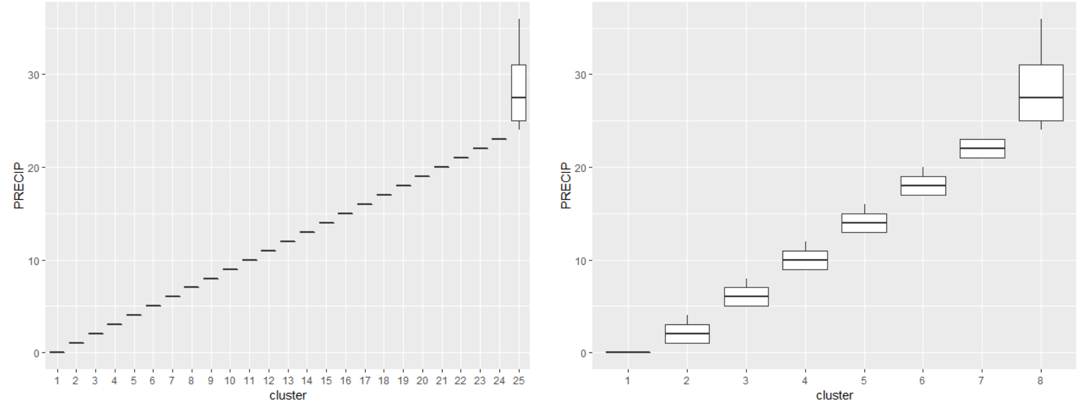
```
> table(PRECIP)

PRECIP
    0     1     2     3     4     5     6     7     8     9    10    11    12
41648   595   362   251   167   140   113    95    75    60    51    42    33
   13    14    15    16    17    18    19    20    21    22    23    24    25
   28    24    17    14    15    14    13    11     9    10     9     4     4
   26    27    28    29    30    31    32    33    34    35    36
    3     3     2     2     2     2     2     1     1     1     1
```

However, by the "table" printed above, there are still too many zero values of $PRECIP$. Therefore we have to develop a method to cluster some values of $PRECIP$ while keeping the clusters significant in terms of predicting response variable.

Here is an algorithm based on the idea of "five-number-summary" to chop up a sorted list of values and regard every part as a cluster:

(1). load the values of $PRECIP$ variable into the algorithm;
(2). sort the values of $PRECIP$ variable;
(3). start from the minimal value loaded into the algorithm;
(4). add one greater value each time;
(5). calculate the 1st quartile Q_1, the 2nd quartile Q_2 and the maximum values max in the algorithm;



(a) Original results of cluster algorithm

(b) Results after merging

Figure 2: Cluster results

(6). if $(max - Q_3) \geq 1.5 \cdot IQR$, then we believe that the added values have an outlier, so ignore the latest added value and regard the remaining ones as a cluster. Throw the cluster out of algorithm. Repeat step (3);
(7). if $(max - Q_3) < 1.5 \cdot IQR$, then we believe that there is no outliers. Repeat step (4).

Finally, by executing the algorithm above, we got 25 clusters as shown in Figure 2(a). And the ANOVA test of those clusters against response variable is conducted.

```
> fit <- aov(`log(1 + pm2.5)` ~ cluster, data = A)
> summary(fit, test = c('Wilks'))

               Df Sum Sq Mean Sq F value   Pr(>F)
cluster         1     30   29.67   29.37 6.03e-08 ***
Residuals   41755  42192    1.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the clusters are significant against response variables, it is not hard to tell that the clusters are too finely-divided. Therefore, just like what we did in the previous section, we merge some clusters while keeping the new clusters still significant.

Table 2: Merging levels of clusters

Old clusters	1	2, 3, 4, 5	6, 7, 8, 9	10, 11, 12, 13	14, 15, 16, 17	18, 19, 20, 21	22, 23, 24	25
New clusters	1	2	3	4	5	6	7	8

The new merged clusters are shown as Figure 2(b). By the ANOVA test,

```
> fit <- aov(`log(1 + pm2.5)` ~ cluster, data = A)
> summary(fit, test = c('Wilks'))

               Df Sum Sq Mean Sq F value  Pr(>F)
cluster         1      9   9.072   8.974 0.00274 **
Residuals   41755  42212   1.011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we know the 8 new clusters are still significant. We simply call the new cluster as variable PRE_clu in further analysis.

2.4 Time, wind, other meteorological factors and $\mathrm{PM}_{2.5}$

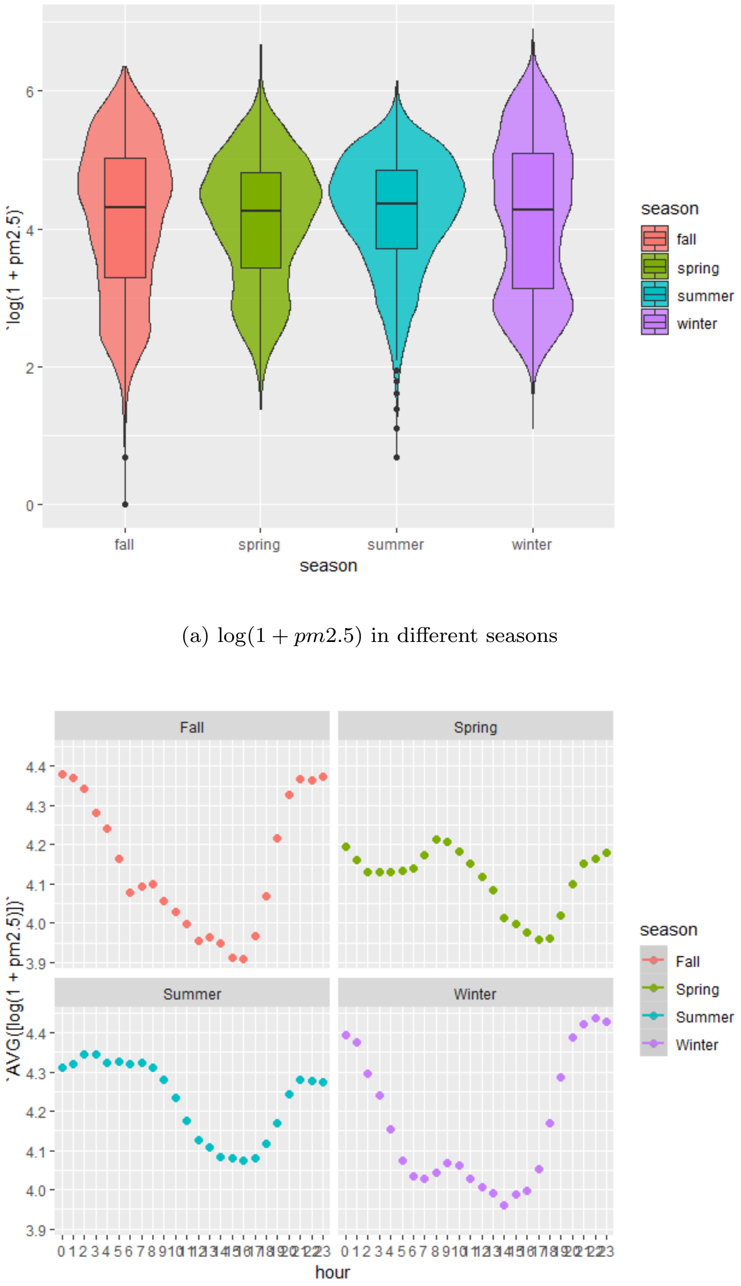
In this part, we would like to explore more about the relationship between $\mathrm{PM}_{2.5}$ and other predictors, which could be divided into several groups such as predictors regarding time, predictors regarding wind, and also predictors regarding other meteorological factors.

2.4.1 Relationship between seasons, hours and $\mathrm{PM}_{2.5}$

Beijing is located in the temperate region of mason climate, enjoying 4 fairly divergent seasons. Both temperature and humidity could influence the concentration of $\mathrm{PM}_{2.5}$ significantly; however, human activity is something we should never ignore. Adjacent to Shanxi Province, which is the biggest and most famous coal-producing region in China, citizens in Beijing always burn a huge amount of coal for heating every winter and therefore pollution could increase dramatically. More information is revealed in figure 3(a).

Next, we will add the factor *hour* (or the merged factor *time*) into our exploration. According to figure 3(b), regardless of the seasons, $\mathrm{PM}_{2.5}$ concentration will fall during 1am to 6am, rise during morning rush hours, fall again in the afternoon, and rise again after 6pm. This specific pattern could be explained by temperature gap between day and night, and also human activity such as traffic jam and residential heating. We notice that in Spring and Summer the peak around morning rush hours is almost at the same level as that after 6pm, whereas in Fall and Winter the peak after 6pm is much higher. This observation gives a hint that residential heating is more polluted than residential cooling in Beijing, and also more polluted than industrial cooling and heating.

In order to give a more quantitive clarification about the difference of $\mathrm{PM}_{2.5}$ between seasons and hours, we use multi-factor ANOVA to test the average gap.



(a) $\log(1 + pm2.5)$ in different seasons

(b) Mean $\log(1 + pm2.5)$ in different seasons and hours

Figure 3: $\log(1 + pm2.5)$ v.s seasons and hours

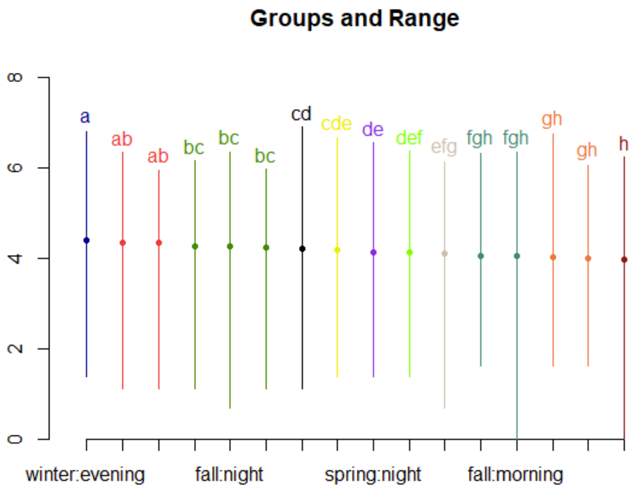


Figure 4: Multiple comparisions: $PM_{2.5}$ against seasons and time

```
> fit <- aov(`log(1 + pm2.5)` ~ season * time, data = dat1)
> summary(fit, test = c('Wilks'))

                Df Sum Sq Mean Sq F value Pr(>F)
season           3     78   26.01   26.13 <2e-16 ***
time             3    420  139.97  140.62 <2e-16 ***
season:time      9    176   19.52   19.61 <2e-16 ***
Residuals    41741  41548    1.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-values of *season*, *time* and the interaction effect are all less than $2 \cdot 10^{-6}$, which is pretty small, we could conclude that the concentration of $PM_{2.5}$ is significantly different in different seasons and hours.

Finally, we could make use of multiple comparisons to explore how $PM_{2.5}$ differs in each time interval. A vivid presentation is given by figure 4, from which we could tell that the pollution is most serious in winter evening, while least serious in spring and fall afternoon.

2.4.2 Relationship between wind and $PM_{2.5}$

In this part, we will explore the influence of variables *cbwd* and $\log(1 + Iws)$ on the pollution level in Beijing.

Take a quick look at the topography of Beijing, we will easily find that there are Yanshan Mountain and Taihang Mountain in the north and west of Beijing, blocking the entry of water vapor; while in the south and east of Beijing, there is a large plain connecting Beijing to the coastal area. By the technique of correlation analysis, from figure 5, we could tell that wind from southeast is more common in Summer while wind from northwest is more common in Winter.

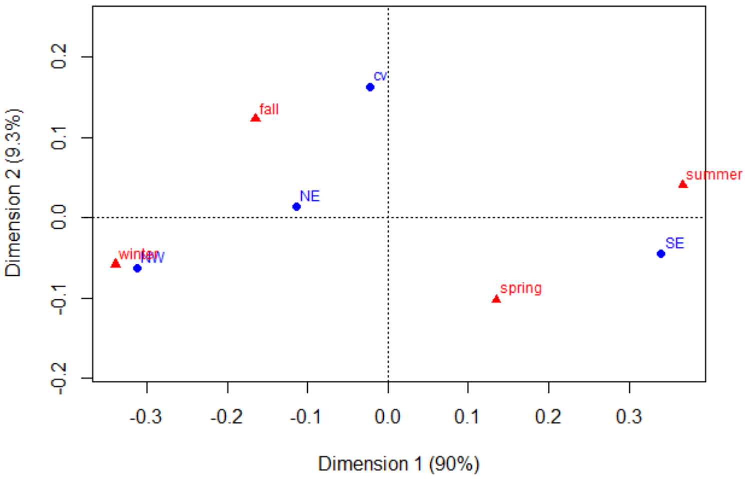


Figure 5: Wind directions in 4 seasons

The chi-square test on cross table also reveals that the relationship is significant, since p-value is extremely small. The information combined gives us a hint that climate in summer could be humid while that in winter could be try.

```
> t <- table(cbdw = dat1$cbwd, season = dat1$season)
> chisq.test(t)

        Pearson's Chi-squared test

data:  t
X-squared = 3408.7, df = 9, p-value < 2.2e-16
```

Next, we will explore more on the duration of wind. Notice that Iws is a cumulative metric of wind speed, and in the scale of time series, it always starts from 0 and keeps increasing until the direction of wind changes. Therefore, by counting the consecutive hours of a certain wind direction, we could generate the time series data of wind duration. From figure 6, we could tell that wind from southeast and northwest is more persistent, and therefore could be more influential.

In the following part, we will focus on the northwest wind and southeast wind. Comparing figure 7(a) and figure 7(b), we could observe a relatively opposite pattern. With northwest wind persisting, $PM_{2.5}$ concentration will first fall and then arise; while with southeast wind persisting, $PM_{2.5}$ concentration will first arise and then fall. It is not easy to come up with a reasonable explanation for this phenomenon, but here is a guess: dry wind is more likely to blow away tiny particles such as $PM_{2.5}$, while wet wind is more likely to bring about fog or smog; however, many factories in heavily polluted industries are located in the north and west of Beijing, while areas in the south and east of Beijing are more clean. These two opposite factors intertwine with each other, and the wet/dry factor works before the polluted particle factor; in other words, if the wind duration is long enough, opposite phenomenon would appear.

In fact, we could also add the wind speed factor into consideration, but it could be combined with analysis of other meteorological factors. Therefore, we would discuss it later.

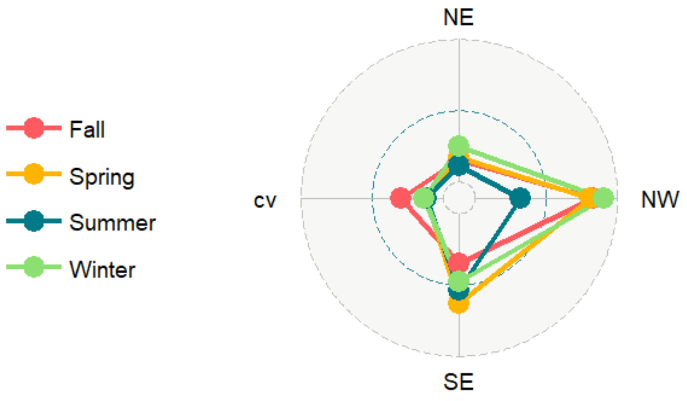
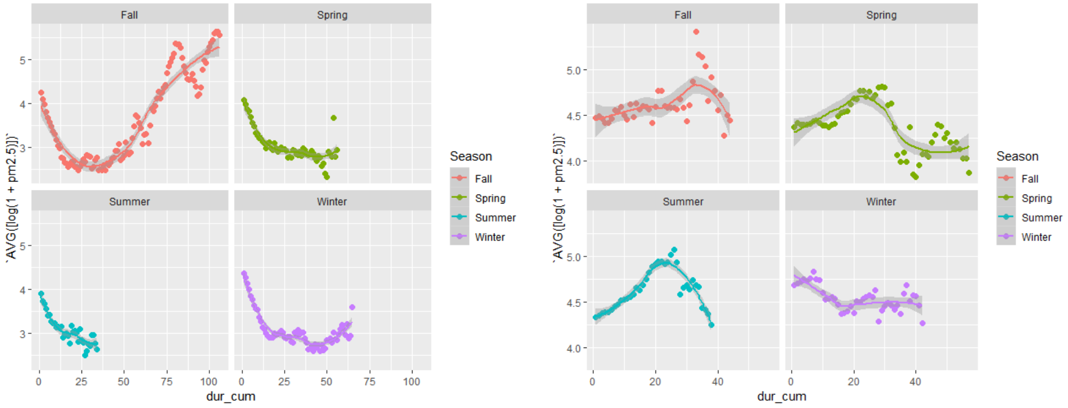


Figure 6: Wind duration of 4 directions



(a) Change in $\log(1+pm2.5)$ with northwest wind persisting

(b) Change in $\log(1+pm2.5)$ with southeast wind persisting

Figure 7: Change in $\log(1+pm2.5)$ with wind persisting

2.4.3 Relationship between precipitation and PM$_{2.5}$

Recall that in section 2.3.3, we created a new categorical variable, PRE_clu, to represent the quantity of rain and snow. In this section, we would like to take advantage of linear regression model to explore the influence of precipitation more specifically.

Notice that, although linear regression model is a powerful tool for prediction, its strong explainability of parameters is also commonly appreciated. Here we do not aim to predict the concentration of PM$_{2.5}$ by using only PRE_clu, but we do hope to get some insight from the parameters of linear models.

Since there are 8 levels for PRE_clu, we have to create 7 dummy variables first:

$$c_i = \begin{cases} 1, & \text{cluster } i \text{ in } PRE\_clu \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1, 2, 3, ..., 7$. Then we build the linear regression model:

$$\log(1 + pm2.5) = \sum_{i=1}^{7} \beta_i \cdot c_i + \epsilon,$$

where $\epsilon \sim N(0, 1)$. We finally got the results as follows.

```
> lm_clu <- lm(log(1 + pm2.5) ~ c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7, data = dat2)
> summary(lm_clu)

Call:
lm(formula = log(1 + pm2.5) ~ c_1 + c_2 + c_3 + c_4 + c_5 + c_6 +
    c_7, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1571 -0.7559  0.1333  0.7628  2.7456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1930     0.1897  16.835  < 2e-16 ***
c_1           0.9641     0.1897   5.082 3.76e-07 ***
c_2           1.2462     0.1916   6.503 7.97e-11 ***
c_3           0.9499     0.1960   4.845 1.27e-06 ***
c_4           0.6612     0.2044   3.234  0.00122 **
c_5           0.6185     0.2204   2.807  0.00501 **
c_6           0.6215     0.2361   2.633  0.00847 **
c_7           0.4122     0.2707   1.523  0.12779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.004 on 41749 degrees of freedom
Multiple R-squared:  0.004049,	Adjusted R-squared:  0.003882
F-statistic: 24.25 on 7 and 41749 DF,  p-value: < 2.2e-16
```

From the output of the algorithm, we could discover that almost all parameters (coefficients) are positive and strongly significant. The concentration of PM$_{2.5}$ will increase if there is precipitation, and with precipitation increasing, the increment will arise first and then fall, reaching almost zero when the cluster is c_8. Although it is anomalous to get positive parameters, we do conclude that a heavy rain/snow will refresh the air, compared with light rain/snow.

2.4.4 Relationship between other meteorological factors and PM$_{2.5}$

Up till now, we have already discussed almost all predictors, including seasons, hours, wind directions, wind duration and precipitation. Those remaining predictors undiscussed are some meteorological factors like Dew Point, Temperature and Pressure, which are all continuous variables. What's more, recall that in section 2.4.2, we simply skipped the analysis of wind speed, so in this part we also take it into consideration.

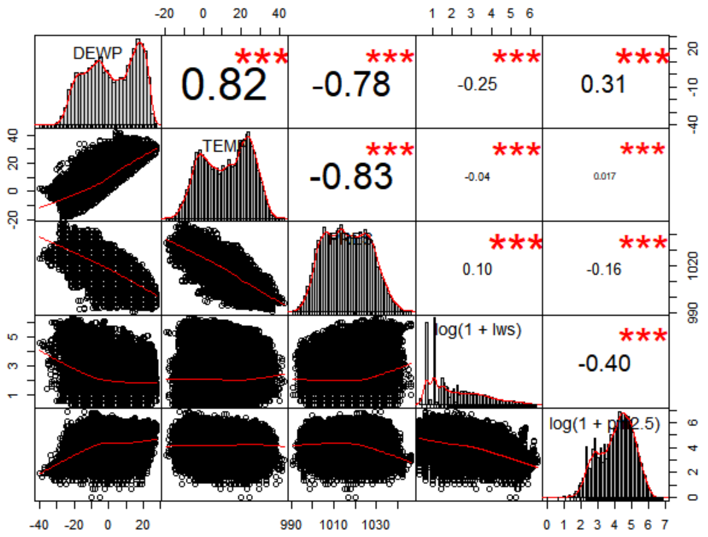


Figure 8: Distribution, correlation, and trend between continuous meteorological variables

From figure 8, a huge amount of information could be revealed. We could observe the distribution of every variable from the diagonal, observe the correlation coefficient between 2 variables on the upper squares off the diagonal (where asterisks represent significance of t-test), and also observe the regression line between 2 variables on the lower squares off the diagonal.

There is no doubt that $DEWP$, $TEMP$ and $PRES$ are highly correlated. Surprisingly, although the correlation coefficients between $\log(1 + pm2.5)$ and other variables are quite small, they are significant in the sense of p-value. It will occur sometimes especially when the total sample size is pretty large.

In general, when the dew point and temperature increase while pressure and wind speed fall, $\text{PM}_{2.5}$ concentration is more likely to increase. Actually, all those variables should be taken into consideration when predicting the level of $\text{PM}_{2.5}$.

3 Machine learning algorithms

In this part, we will try to predict $\log(1 + pm2.5)$ using machine learning algorithms based on tree models. Tree models are powerful for regression and prediction, and enjoy good interpretation as well.

3.1 Decision Trees

Decision Trees can be quite simple and of low accuracy, but the variables each node picks are good hints for interpretation. We first split the dataset into training set (around 60%) and test set (around 40%), and then input $\log(1 + pm2.5)$ as response variable, and $DEWP$, $TEMP$, $PRES$, $\log(1 + Iws)$, *time*, *season*, *cbwd* and PRE_clu as predictors.

```
> tree_dat3 <- rpart(log_pm2.5 ~ DEWP + TEMP + PRES + log_Iws +
+                      time + season + cbwd + PRE_clu,
+                    data = dat3, subset = tindex, method="anova")
> printcp(tree_dat3)

Regression tree:
rpart(formula = log_pm2.5 ~ DEWP + TEMP + PRES + log_Iws + time +
```

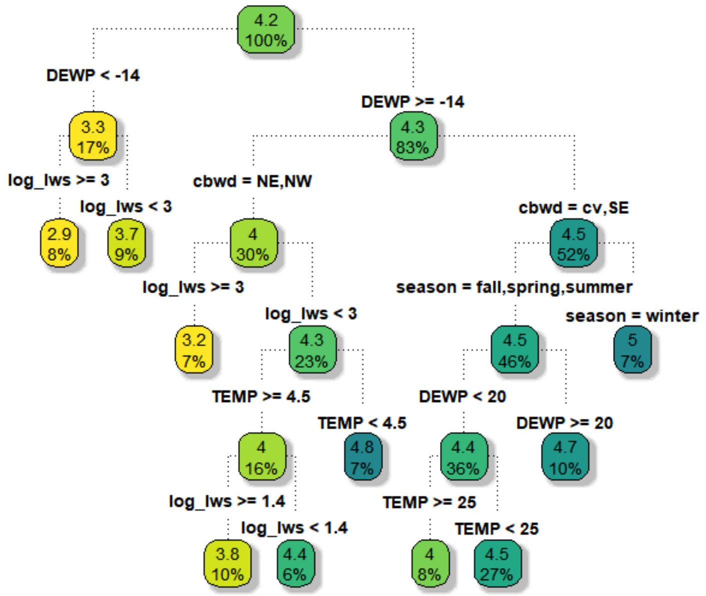


Figure 9: The decision tree

```
    season + cbwd + PRE_clu, data = dat3, subset = tindex, method = "anova")

Variables actually used in tree construction:
[1] cbwd    DEWP    log_Iws season  TEMP

Root node error: 25285/25054 = 1.0092

n= 25054

        CP nsplit rel error  xerror      xstd
1 0.147346      0   1.00000 1.00004 0.0073611
2 0.058874      1   0.85265 0.85354 0.0073522
3 0.029701      3   0.73491 0.73089 0.0066410
4 0.027512      4   0.70521 0.70714 0.0065734
5 0.019405      5   0.67769 0.67797 0.0063907
6 0.013295      6   0.65829 0.66095 0.0062352
7 0.010139      7   0.64499 0.64834 0.0061730
8 0.010000      9   0.62471 0.63613 0.0060681
```

The algorithm chose wind direction, dew point, cumulated wind speed, season and temperature as predictors. There are 9 internal nodes in the decision tree, with both decreasing training error and decreasing test error. Therefore, we do not need to prune the decision tree since more internal nodes will not lead to overfitting. From figure 9, we could tell that dew point is the first thing to consider, and then are wind speed, wind direction, season and temperature. Internal nodes reveal the specific order of decisions, while terminate nodes give the ultimate answers.

However, the test error rate is $1.0092 \cdot 0.63613 \cdot 100\% = 64.198\%$, which is too high. The reason could be the inefficiency of a single tree when predicting, and also the lack of scaling. Therefore, we will try random forest algorithm and scale the dataset in the following part, in order to increase the performance of prediction.

3.2 Random forest

Notice that although scaling could be powerful, we have to keep the response variable as original for the sake of prediction.

We construct 100 decision trees for this random forest to achieve good performance. What's more, for each bagged tree in this forest, only $\lceil p/3 \rceil = \lceil 8/3 \rceil = 3$ variables are chosen as split candidates from the full set of $p = 8$ predictors to decorrelate the trees.

Finally, we got the list of features ordered by importance:

```
> set.seed(1)
> rf_dat3 <- randomForest(log_pm2.5 ~ DEWP + TEMP + PRES + log_Iws +
+                             time + season + cbwd + PRE_clu,
+                         data = dat3_scale, subset = tindex,
+                         ntree = 100, mtry = 3, importance = T)
> I <- importance(rf_dat3)
> I <- I[order(I[,2], decreasing = T),]
> print(I)
```

```
         %IncMSE IncNodePurity
DEWP    90.63622     6241.7367
log_Iws 76.37524     4869.9631
TEMP    64.87924     3460.2998
cbwd    90.16017     2786.0048
PRES    38.42738     2762.1500
season  39.97748     1722.4040
time    46.82376     1001.3598
PRE_clu 16.83790      257.0795
```

which is quite similar to the order of features in the decision tree. Surprisingly, dew point is not only the very first feature to consider when it comes to predict $PM_{2.5}$, but also the most important feature. Wind speed, wind directions, and temperature are also quite important. However, precipitation is not that influential compared with other features.

At last, we could calculate the test error (percentage) of random forest algorithm by the formula:

$$error = \frac{\sum(\hat{y_i} - y_i)^2}{\sum(y_i - \bar{y})}.$$

```
> ytrue <- dat3$log_pm2.5[-tindex]
> yhat <- predict(rf_dat3, newdata = dat3_scale[-tindex, ])
> mean((yhat - ytrue)^2) / mean((ytrue - mean(ytrue))^2)
```

```
[1] 0.3559027
```

Compared with the error of 64.198% for a single decision tree, random forest generates error of only 35.590%, which is not only a huge progress of performance, but also the accuracy we are satisfied with. Therefore, we could use the random forest model already built to impute the missing values of $PM_{2.5}$, and give a basic judgement whether the air quality in Beijing had improved or not.

4 Projection

4.1 Imputation of missing $PM_{2.5}$

In this part, we will impute the missing values of $PM_{2.5}$ using the random forest model we have built. The specific steps could be described as follows:

(1). Assign values of PRE_clu according to the results we got in section 2.3.3.

(2). Create a new data set $dat.p$ including columns $year$, $month$, day, $hour$, $DEWP$, $PRES$, $\log(1 + Iws)$, $time$, $season$, $cbwd$, PRE_clu and $\log(1 + pm2.5)$ (with missing values).

(3). Scale the new dataset $dat.p$ into another new dataset $dat.p\_scale$ but keep the response variable $\log(1 + pm2.5)$ as original.

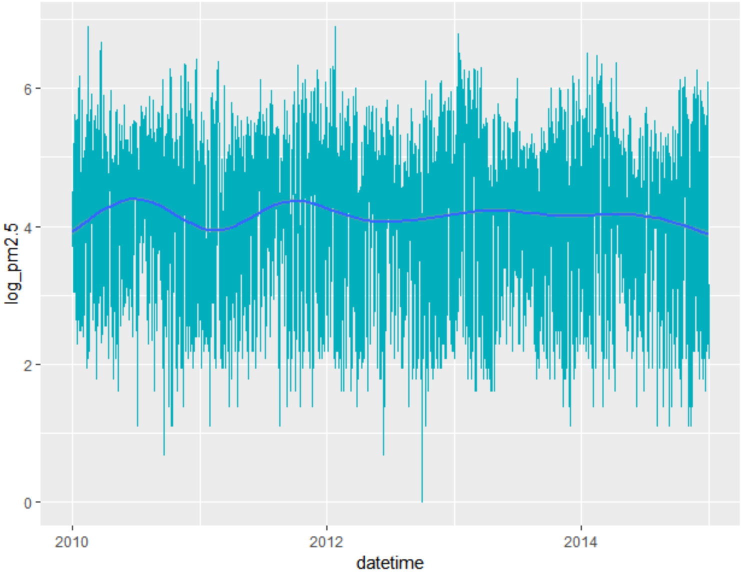


Figure 10: Time series of $\log(1+pm2.5)$

(4). Split the dataset $dat.p$ into two datasets, one of which is called $dat.p\_FULL$ and is made of complete cases; the other one is called $dat.p\_NA$ and is made of cases with missing values. At the same time, $dat.p\_scale$ is also splited into two datasets, $dat.p\_scale\_FULL$ and $dat.p\_scale\_NA$.

(5). Predict the variable $\log(1+pm2.5)$ in the dataset $dat.p\_NA$, by fitting the dataset $dat.p\_scale\_NA$ into the random forest model built previously.

(6). Merge all cases in datasets $dat.p\_FULL$ and $dat.p\_NA$, and order the cases by date and time.

4.2 Trend of PM$_{2.5}$ from 2010 to 2014

At the end, we got the time series data $\log(1+pm2.5)$ without missing values. From figure 10, it could be hard to tell the specific trend since it looks quite stable. However, a slight decreasing pattern could be observed if any.

In fact, we do have non-parametric methods to test and analyze the specific trend. Here we will make use of Cox-Stuart test in order to get a basic judgement. Cox-Stuart test splits the time series data into the first half and the second half, pairs values from the two halves, counts the signs of differences of those pairs, and finally calculates the p-value based on binomial distribution.

```
> library(randtests)
> cox.stuart.test(ts$log_pm2.5, 'two.sided')

        Cox Stuart test

data:  ts$log_pm2.5
statistic = 10675, n = 21792, p-value = 0.002813
alternative hypothesis: non randomness

> cox.stuart.test(ts$log_pm2.5, 'left.sided')

        Cox Stuart test

data:  ts$log_pm2.5
statistic = 10675, n = 21792, p-value = 0.001406
alternative hypothesis: decreasing trend
```

```
> cox.stuart.test(ts$log_pm2.5, 'right.sided')

        Cox Stuart test

data:  ts$log_pm2.5
statistic = 10675, n = 21792, p-value = 0.9987
alternative hypothesis: increasing trend
```

From the test results above, we could decide that there is a significant decreasing trend of $\log(1 + pm2.5)$, though it is not that obvious from the figure 10. The decreasing pattern is a good indicator for the improvement of air quality in Beijing, and therefore encourages citizens to do more to protect the environment.

5 Conclusion

From the analysis above, we could reach abundant information and conclusions. To be more specific, $\mathrm{PM}_{2.5}$ is higher in Winter, rush hours and night; northwest wind ultimately increases concentration of $\mathrm{PM}_{2.5}$ while southeast wind ultimately decreases it. However, when it comes to predict the concentration of $\mathrm{PM}_{2.5}$, dew point, wind speed and temperature are more influential.

Although the concentration of $\mathrm{PM}_{2.5}$ enjoys a slightly decreasing trend from 2010 to 2014, many problems are still needed to be solved. The geographical layout of the factories around Beijing should be adjusted, especially those factories with serious pollution; and citizens should also use less coal for heating and cooling, and use more clean energy such as natural gas. With actions and efforts, we believe that air quality in Beijing must get better in the future.

References

[1] Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.