**Exploration**

Our dataset comes from the U.S. Department of State in the Embassy and Consulates of five major cities in China. Data is measured in terms of the total mass of particulate matter (particles with size 2.5 micrometers or less, which we denote as "PM 2.5"). The PM 2.5 data is hourly over the course of 2015, so there are 24 data points in each day for each of the five cities.

The nature of this dataset (and personal experience with air pollution) led us to a hypothesis that the pattern of P.M 2.5 varies from hour to hour. So we focused on four specific hours in the first stage of our exploration: We collected data at 12:00am, 8:00 am, 12:00 pm, and 6:00pm. 8:00 am and 6:00 pm are chosen since they are rush hours where we expect to see primary levels of transportation. 12:00am seems to be a solid baseline for low human activity and and 12:00 pm is the midday, where we might expect lower transportation but higher industrial air pollution.

Based on an informal comparison of these four hours, we made the assumption that the P.M 2.5 would be highly correlated with air pollution from traffic. We therefore expected to see relatively low P.M 2.5 at 12:00am and relatively high values at the 8:00 am and 6:00pm measurements.

Below in Figure 1 is our PM 2.5 plot for all five cities at the four hours. Note that the graph is color-coded by site: SH means Shanghai, GZ is Guangzhou, CD indicates Chengdu, BJ is Beijing, and SY stands for Shenyang.



Figure 1

Each plot is the value of everyday PM2.5 at a particular time point (12:00 am, 8:00 am, 12:00 pm, and 6:00 pm) in 5 big cities in China (Shanghai, Guangzhou, Chengdu, Beijing, and Shenyang).

Despite having removed missing values, the deensity of points from these five cities makes it hard to notice any pattern. So, as an exploration, we decided to focus the first 50 days of 2015 which gives us the following graph.



Figure\*.3

Four hours plots of 5 cities for the first 50 days

In order to make the pattern for each city more clearly we concentrate on Hour 0 and split cities apart.



Figure\*.4

Hour 0 plots of 5 cities for the first 50 days

Figure 4 indicates that for the given hour and day period, Shanghai and Guangzhou maintain a very stable pattern. The P.M. 2.5 for Chengdu fluctuate a little bit. However, its pattern keeps stable in general. For Beijing and Shenyang, however, their P.M. 2.5 change violently for the first 50 days. Different latitudes and longitudes mays provide some information for us. Shanghai and Guangzhou are in the south of China. However, Beijing and Shenyang are in the north of China.

There is a very clear trend to go up for Beijing. It seems that the P.M. 2.5 for Beijing will be higher than other cities on average for the following days. This very intuitive analysis gives us a direction that we should focus our project on Beijing which is also the reason why we are trying to use the meteorological data to explain the P.M. 2.5 for Beijing in the second part of our project.

So in the following part, two graphs for the air condition of these five cities in days 51-100 and days 101-150 will be plotted first. As mentioned above, we expected to see that Beijing has higher P.M. 2.5 than the rest of cities. We also want to compare the different patterns from different hours.



Figure 5

Four-hour plots of 5 cities for days 51-100



Figure 6

Four-hour plots of 5 cities for days 101-150

Figure 5 shows that Beijing has higher PM 2.5 values for the following 50 days. However, this pattern isn’t significant. In fact, Shenyang exceeds Beijing for a period of time. Both of the two plots indicates that PM 2.5 patterns of Beijing does change via hour to hour. However getting an accurate summary based on these plots is really hard. So in the next part we turns to regression model explaining Beijing’s P.M. 2.5 via meteorological data.

Outliers

As we can see in the previous plot, the outliers of 2015 were concentrated in January, Feburary, October and December. These are the months when temperatures are relatively low within a year. Naturally, we want to see if this phenomenon is a pattern across 3 years.



Figure1

Figure1: each color represents mean pm value of every day in one year. 2013 was blue, 2014 was green, 2015 was black. The red line is horizontal line with value 250.

As we can tell from figure 1, the points that exceeded the red line which considered as outliers, mostly were concentrated in the months of January, Feburary, March, October and December. The exact day are shown in table 1.

Table1: Days of PM2.5 values over 250 for 3 years.



In the regression part, there is a linear relationship between the pm2.5 values excluding extreme values and meteorological variables. So for the outliers, I also wanted to check if there is a linear relationship as before. The fitted results are plotted in figure 2.



Figure 2

Figure 2: Fitted values of a linear regression for outliers in 2015.

There was only one meteorological variable is significant, that is mean visibility measured in miles. This variable was kind intuitively reasonable, however, maybe too directly dominant the relationship. So I refitted the model without this variable, it turned out that the mean sea level pressure became significant.

When I could not find a reason with my knowledge to explain the relationship, I decided to see if there is a similar relationship in one particular day. So I took 24 Pm2.5 values of January 15 in 2015, did regression on the corresponding meteorological variables within a day. The regression result was shown in Figure 3. It turned out that again the sea level pressure was significant with adjusted R-square as highly as 0.77.



Figure 3

Figure 3: Fitted values of a linear regression for PM2.5 values in January 15th,2015.

To be continued: examine more outlier days, and explain the significance of sea level pressure variable.

**Chapter II Regression fitting and Outlier Analysis**

Part I:Regression of meteorological factors on daily average PM2.5

In the second part of our project, the key point is to figure out the relationship between meteorological variables and PM2.5 values in each day during 2015. Based on the analysis above about the hourly PM2.5 for 5 cities, Beijing is outstanding than other cities, presenting a higher level as a whole. Thus, our first target for study is Beijing, the capital as well as political center of China.

The meteorological data we obtained is from:

<http://www.wunderground.com/history/>.

In accordance with the daily average value of PM2.5, the meteorological factors being analyzed are the mean values of temperature, dew point, humidity, sea level pressure and wind speed in each day from January 1 to December 31 of 2015. Taken the lag effect into consideration, a lag factor is also included in our regression models, which is the average PM2.5 value one day before.

First we combine the meteorological data with the daily averagePM2.5 data, then to fit our regression model. However, after drawing a graph of daily average PM2.5 alone, what we noticed is that it has a high beginning and end. In order to exclude the impacts of these extremely high values on our final model, it is obvious that we should focus only on days with normal PM2.5. According to the new air quality standards in China, we delete days that have average PM2.5 higher than 250μg/m³, which represents for a severe air pollution.



Figure 1

Figure 1 shows the daily average PM2.5 of Beijing during 2015. The red line stands for the value equal 250. Points above the red lines are days we delete in our model fitting.

Next step, we start with the linear regression without interaction terms after filtering out days with PM2.5 larger than 250μg/m³. And for days PM2.5>250μg/m³, specific analysis is needed for further research.

By summarizing the fitted model with all meteorological variables and one lag factor, we find out that only four variables are significant, which are the daily average values of temperature, sea level pressure, wind speed and the lag factor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimated | Std.Error | t value | Pr(>t) |
| Intercept | 1561.10430 | 409.77912 | 3.810 | 0.000165 \*\*\* |
| Lag(mean.pm,1) | 0.44853 | 0.03891 | 11.527 | < 2e-16 \*\*\* |
| Mean.temperature | -1.20870 | 0.21537 | -5.612 | 4.14e-08 \*\*\* |
| Mean.sealevelpressure | -46.72900 | 13.27806 | -3.519 | 0.000491 \*\*\* |
| Mean.windspeed | -8.56000 | 0.69009 | -12.404 | < 2e-16 \*\*\* |

Table 1

The table contains all the significant variables for the linear model.

Based on the table above, the estimated coefficients of mean temperature, mean sea level pressure and mean wind speed for each day are negative, which indicates the negative effects of these meteorological variables on the PM2.5 values. In contrast, the positive coefficient of lag(mean pm,1) suggests a positive relationship between pm2.5 and the lag factor.

Then we need to determine how well the model is. However, the adjusted R-squared is 0.5116, indicates the fitted model is not ideal. And in conjunction with the residual plot, which shows a strong pattern, we can say that a

Non-linear regression is necessary in our future exploration.



Figure

Figure 2 is the residual plot of linear regression on significant meteorological factors versusPM2.5



Figure

Figure 3 is the fitted model plot of linear regression PM2.5 versus Meteorological factors.