Creating a pollution warning system:

A technical report on predicting degree of air pollution in Guangzhou, China

University of Illinois at Urbana-Champaign

Group 2

Introduction

In the recent years, China has been experiencing an increasing level of air pollution across its major cities. Guangzhou, a major southern city in China, collected hourly air pollution throughout year 2012 to 2015 along with data on other environmental factors. In this dataset, PM2.5 value is collected as the indicator of air pollution level, as it represents the concentration of fine particles in the air. In this report, we attempted to create a statistical model to predict air pollution level for the city based on previously collected data. Based on our prediction model, we will create a daily alert system to inform citizens of Guangzhou on whether or not they should wear a mask outside. Additionally, for the purpose of educating the public, we will generate daily, weekly, and monthly summary information for air pollution trends in Guangzhou.

Data Description

The data is hourly PM2.5 values collected from three sites of Guangzhou from 2011 to 2015. It contains 52584 observations and 17 variables, including 7 categorical variables which are *numbers, year, month, day, hour, season, cbwd* and 11 continuous variables *which are PM from 3 sites, DEWP, TEMP, HUMI, PRES, Iws, precipitation and Iprec*. The description of the variables is listed below:

Variables	Functions
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

season	Season of data in this row
PM_City Station	PM2.5 Concentration(ug/m^3) from City Station
PM_5th Middle School	PM2.5 Concentration(ug/m^3) from 5th Middle School
PM_US Post	PM2.5 Concentration(ug/m^3) from US Post
DEWP	Dew Point(Celsius Degree)
TEMP	Temperature(Celsius Degree)
HUMI	Humidity(%)
PRES	Pressure(hPa)
cbwd	Combined wind direction
Iws	Accumulated wind speed(m/s)
precipitation	Hourly precipitation(mm)
Iprec	Accumulated precipitation(mm)

The descriptive statistics of variables are listed below:

Variable	Minimum	Mean	Maximum
DEWP	-11.70	17.34	27.40
HUMI	13.00	78.14	100.00
PRES	975.00	1004.90	1023.10
TEMP	1.70	21.76	37.40
Iws	0.00	8.08	214.30
precipitation	0.00	0.24	90.40
Iprec	0.00	1.01	126.00
Average PM2.5	1.00	52.15	327.00

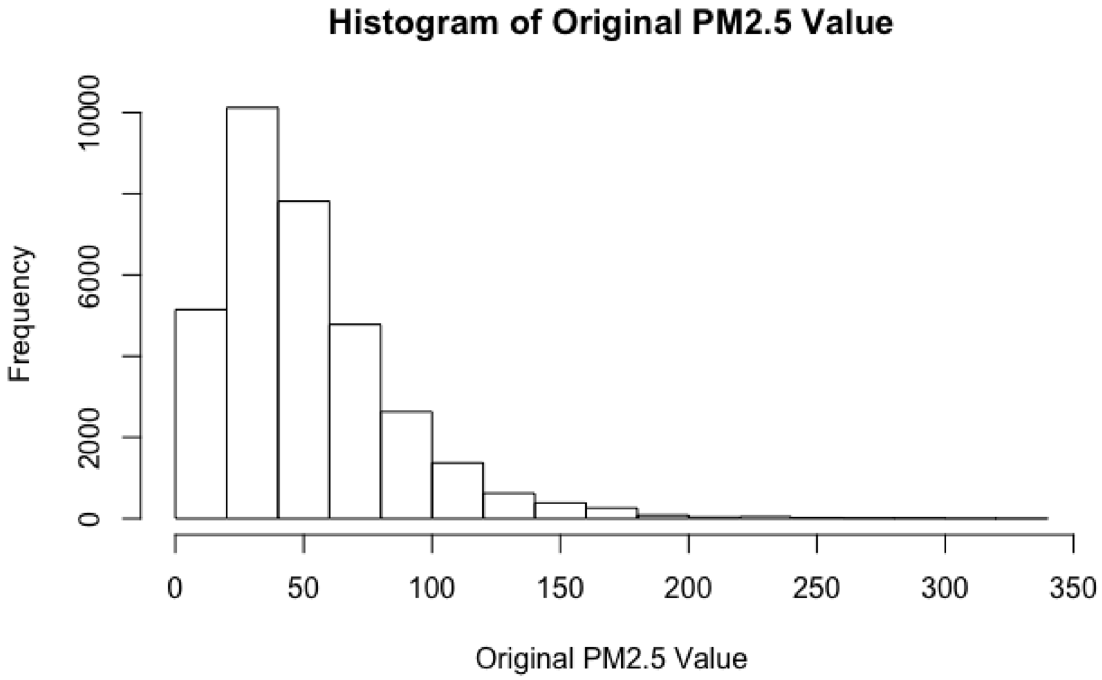
Continuous Variables

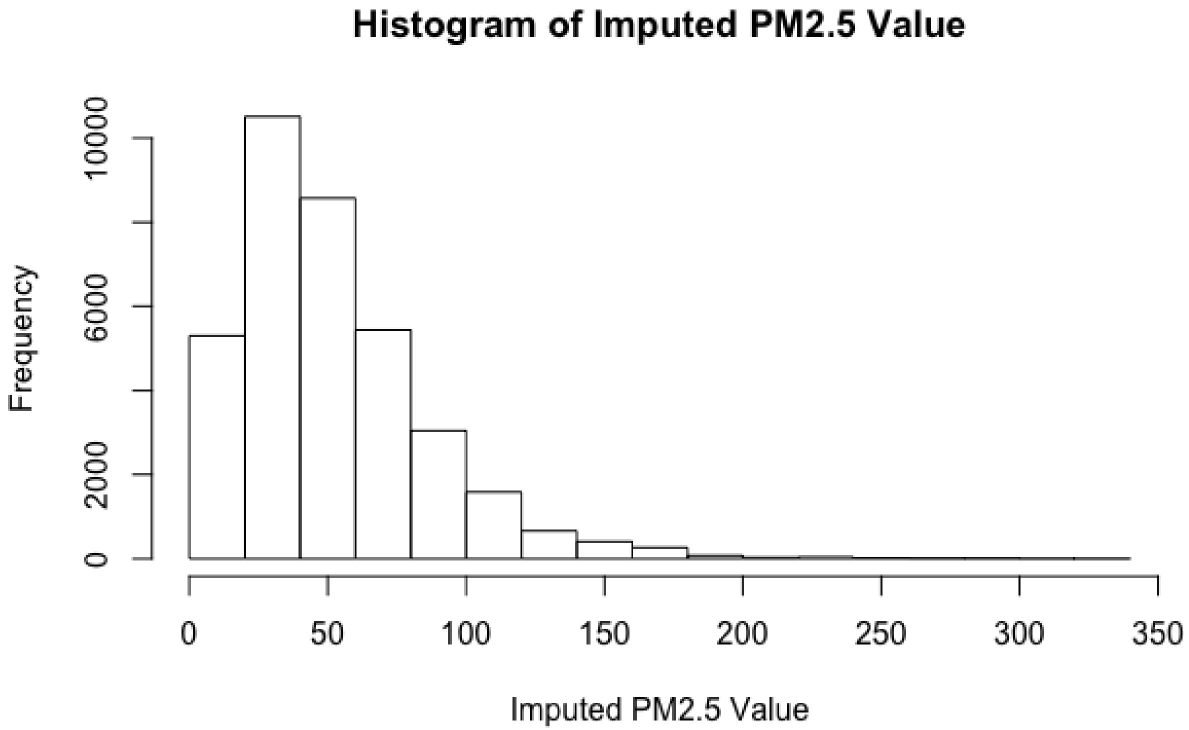
Variable	Levels
Year	2011-2015
Month	1-12
Day	1-31
Hour	0-23
Season	1-4
cbwd	cv, NE, NW, SE, SW

Categorical Variables

Data Cleaning

Upon receiving the data, we researched previous research papers that had analyzed this data and any background knowledge required to understand our variables. There are several problems embedded in our dataset: 1. Two of our PM2.5 collection points have identical measurements (City Station and US Post) 2. There are continuous blocks of missing data for middle school station from year 2010 and year 2011. There are random missing values scattered in our dataset. To solve the first problem, we simply ignored one of the two overlapping data collection points, because including both won’t provide us extra information on the PM2.5 value of the day. To fix the second problem, we decided to remove the data from year 2010 to late 2011. 3. During the data cleaning, we found there are some extremely high PM2.5 value such as 940 and 937 which are obviously not consistent with the values in the rest hours of a day. Therefore, we decided to delete these data directly. 4. For the third problem, we imputed data using K Nearest Neighbor (KNN) method to fill out the missing PM2.5 values from year 2012 to year 2015. We chose KNN method because we assume that any missing hourly PM2.5 value will follow the general trend of adjacent hourly PM2.5 values in the same day. After imputing missing values, we created histograms to present data distributions of both the original dataset and our imputed new dataset. We examined the two data distributions to determine whether or not our imputed data values follow the original data distribution. Because the two data distributions look approximately the same, we can say with confidence that our imputed data values follow the original data distribution and are valid to be included in our analysis. The two data distribution histograms are provided below:





Model Selection

Before selecting a prediction model, we need to clarify our purpose for the prediction in order to compare the effectiveness of different models. According to our client's requests, our goal should be creating a warning system to inform the public about potential median polluted days (PM2.5>75) and potential severely polluted days (PM2.5>150). Therefore, we will perform classification using various prediction methods and use the classification accuracy to identify our best model.

The prediction methods we tested and compared are: linear regression, decision tree, and random forest. We used the average PM2.5 values between the two stations as our response variable. And we chose all other environmental factors excluding *season* and *precipitation* as predicting variables by automatic variable selection statistical method. We also performed train-split on our dataset based on randomly selected days. The train-split process let us train our prediction models on 70% of our original dataset and test the effectiveness of our models on 30% of our data. For each model we generated, we predicted each day's hourly PM2.5 values based on the corresponding PM2.5 values and other environmental factors from the day before. After obtaining the predicted hourly values, we chose the maximum PM2.5 values within each day and classified these values into three groups: 1) PM2.5<75, 2) 75<PM2.5<150 and 3)150<PM2.5. Then, we classified the original daily maximum PM2.5 to the same three groups and compare them with our prediction groups. To make the comparison more comprehensible, we created classification accuracy tables to show each model's performance.

Linear Regression	**Pred Group 1 (PM2.5<75)**	**Pred Group 2 (75<PM2.5<150)**	**Pred Group 3 (150<PM2.5)**
Actual Group 1 (PM2.5<75)	212	22	0
Actual Group 2 (75<PM2.5<150)	80	96	2
Actual Group 3 (150<PM2.5)	6	31	1

Tree	**Pred Group 1 (PM2.5<75)**	**Pred Group 2 (75<PM2.5<150)**	**Pred Group 3 (150<PM2.5)**
Actual Group 1 (PM2.5<75)	223	11	0
Actual Group 2 (75<PM2.5<150)	109	69	0
Actual Group 3 (150<PM2.5)	16	21	1

Random Forest	**Pred Group 1 (PM2.5<75)**	**Pred Group 2 (75<PM2.5<150)**	**Pred Group 3 (150<PM2.5)**
Actual Group 1 (PM2.5<75)	218	16	0
Actual Group 2 (75<PM2.5<150)	99	79	0
Actual Group 3 (150<PM2.5)	15	23	0

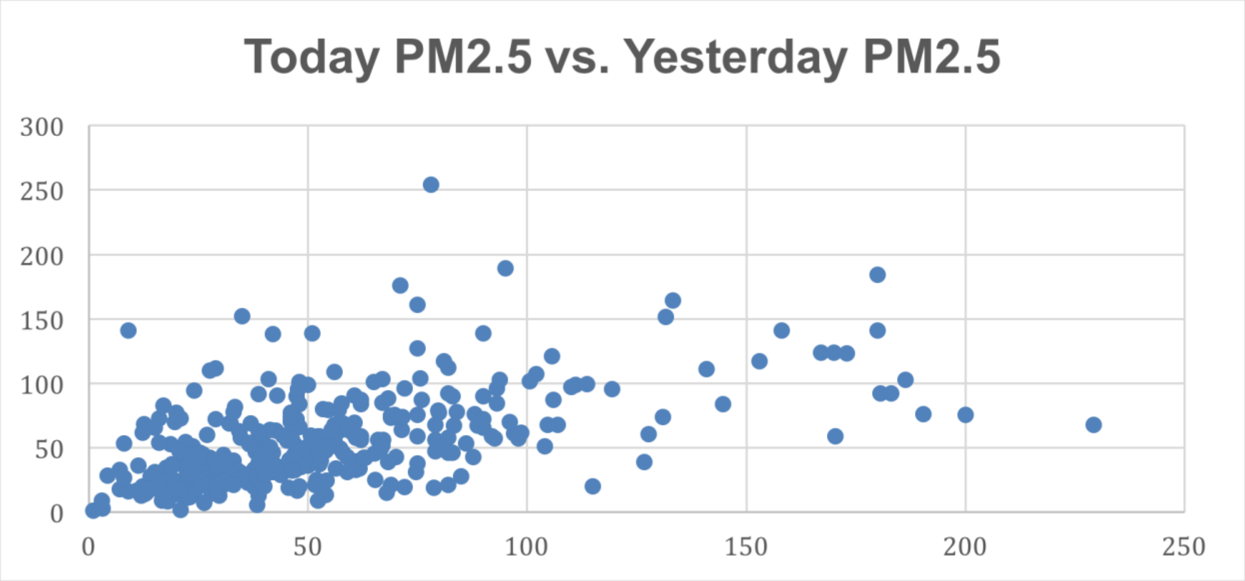
To compare the classification performance of our three models, we calculated four percentages based on each of their accuracy tables to reflect their prediction effectiveness: 1) percentage of correct grouping, where the daily PM.2.5 in predicted and original data values fall into the same classification group. 2) Percentage of redundant warning, where actual group 1 PM2.5 values match to predicted PM2.5 values in group 2 or 3. In other words, the predicted PM2.5 value for that certain day is much higher than the actual PM2.5 value of the day and we will give out an unnecessary warning. 3) Warning error, where actual group 2 PM2.5 values (75< PM2.5 <150) match to predicted PM2.5 values in group 1; or actual group 3 PM2.5 values (150<) match to predicted PM2.5 values in group 2. In other words, the predicted PM2.5 value for that certain day is much lower than the actual PM2.5 value of the day. 4) Adjusted accuracy, where we count actual group 3 PM2.5 values that were predicted to be group 2 PM2.5 values as correct prediction. We calculate this adjusted accuracy because we would still be able to warn the public to some extent in these situations.

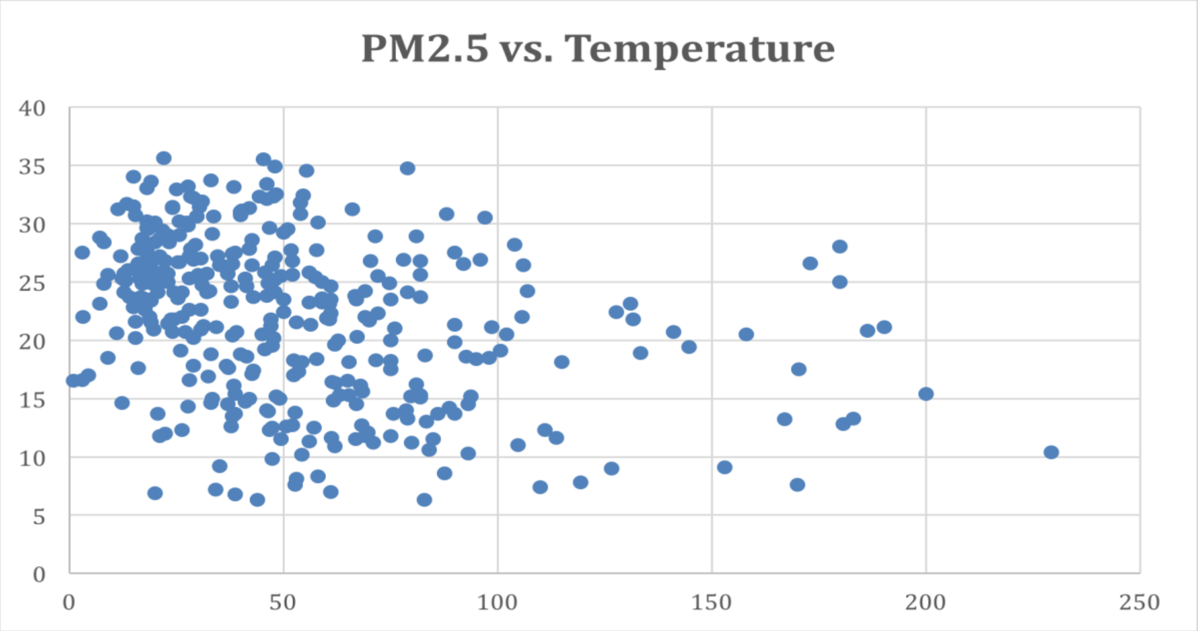
Model Type	Redundant	Accuracy	Adjusted Accuracy	Warning Error
Linear Regression	0.05	0.69	0.76	0.19
Tree	0.02	0.65	0.70	0.28
Random Forest	0.04	0.66	0.71	0.25

To select the best prediction and classification model, we look for the model that gives the highest true and adjusted classification accuracy with the lowest classification warning error. Therefore, we select linear regression as our best prediction model to perform classification and create the three predicted PM2.5 value groups.

Further Analysis

Even though linear regression is our best prediction model, its accuracy and warning error are still not ideal. We plotted some of our dataset to explore reasons behind this problem:





Looking at these two scatter plots, we can see that data points for PM2.5 values above 150 are very scarce and spread out. Indeed, we found that only 2% of the data ave PM2.5 values above 150. In this case, it would be impossible for our linear regression model to capture the behaviors of these data points, which had resulted in the poor classification accuracy for the group3 data.

Our group was not able to find an effective way to solve this problem, but we will find a creative way to achieve our goal of creating a capable warning system.

Warning System

We believe that our predicted PM2.5 values generally fall short of the actual PM2.5 values. Therefore, to classify our predicted values into the same groups as the actual values, we artificially lower the PM2.5 value cutoffs for both groups: if the predicted PM2.5 value is lower than 60, we classify it as group 1; if the predicted PM2.5 value is between 60 and 130, we classify it as group 2; and if the predicted PM2.5 value is larger than 130, we classify it as group3. After these adjustment, we can clearly see improvements in our classification from the model classification and performance table below:

Linear Regression	Pred Type 1	Pred Type 2	Pred Type3
Type 1	176	58	0
Type 2	31	143	4
Type 3	2	35	1

Model Type	Redundant	Accuracy	Adjusted Accuracy	Warning Error
Linear Regression	0.14	0.71	0.79	0.07

Based on this adjusted classification method, we can create an air pollution warning system for the city of Guangzhou. Each day's environmental factors can be used to predict the maximum PM2.5 values for the next day. We will be able to classify these predicted PM2.5 into three groups that correspond to different levels of air pollution: group1 is light air pollution, which means the citizens of Guangzhou will not need to wear a mask outside; group2 is medium air pollution, which we suggest citizens of Guangzhou to wear a mask outside; group3 is high air pollution, which we suggest the citizens to stay inside. We will be able to give proper warnings to the public with 79% accuracy. The downside of our warning system is that we will give unnecessary warnings 14% of the time and we won't be able to give the public any warnings about 7% of the time. However, we are not too concerned about these unnecessary warnings, because telling people to wear masks during slightly polluted days will not harm their health. Another restriction of our warning system is that, we won't be able to detect group 3 air pollution levels very well. This results from the small amount of extremely high PM2.5 in our dataset. However, we will be able to predict and classify most highly polluted days into group 2, which will still warn the citizens to wear a mask outside.

Public Education

To better understand the general trends of PM2.5 values in Guangzhou, we created data visualizations based on various time intervals: hourly trend in a day, daily value in a week, weekly trend in a month, monthly trend in a year and seasonal trend in a year. Here we interpret these visualizations to provide general information on PM2.5 values for the public.

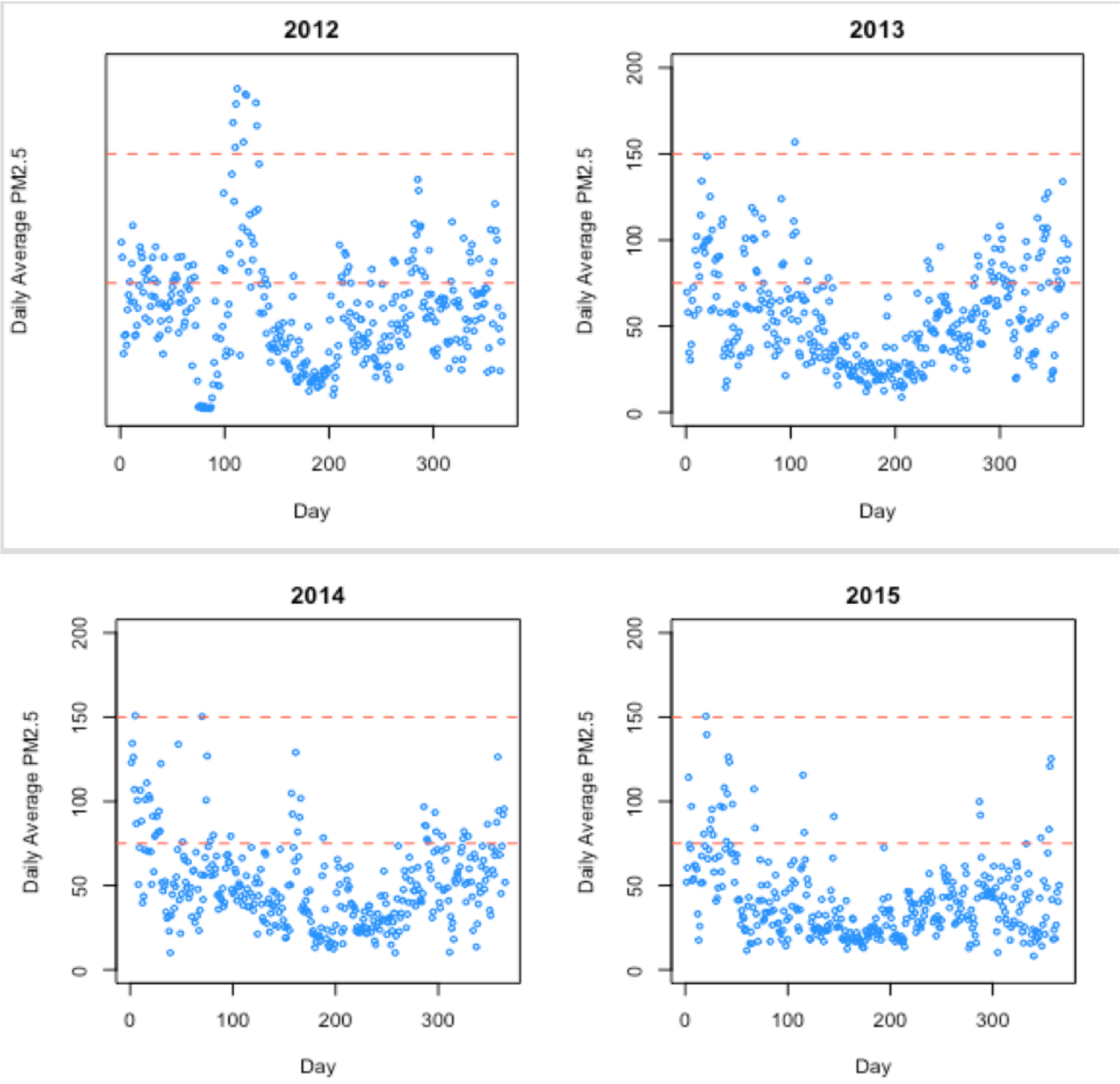
Data Visualizations and Interpretations

1. Trend over years and within a single year

We plotted the daily average PM2.5 for the continuous four years to observe if there is any particular trend over time. Based on Figure 1, it seems that the U-shape patterns are similar during the most recent three years which suggests that the air quality is the best during the middle of a year and the air is most severely polluted in the beginning and end of a year. While in the spring of 2012, the average PM2.5 are higher than any other period during the four years.

Here we suspect some extreme events happened in the spring of 2012 that caused the dramatic increase of air pollution level.

Figure 1. Guangzhou Daily Average PM2.5, 2012-2015

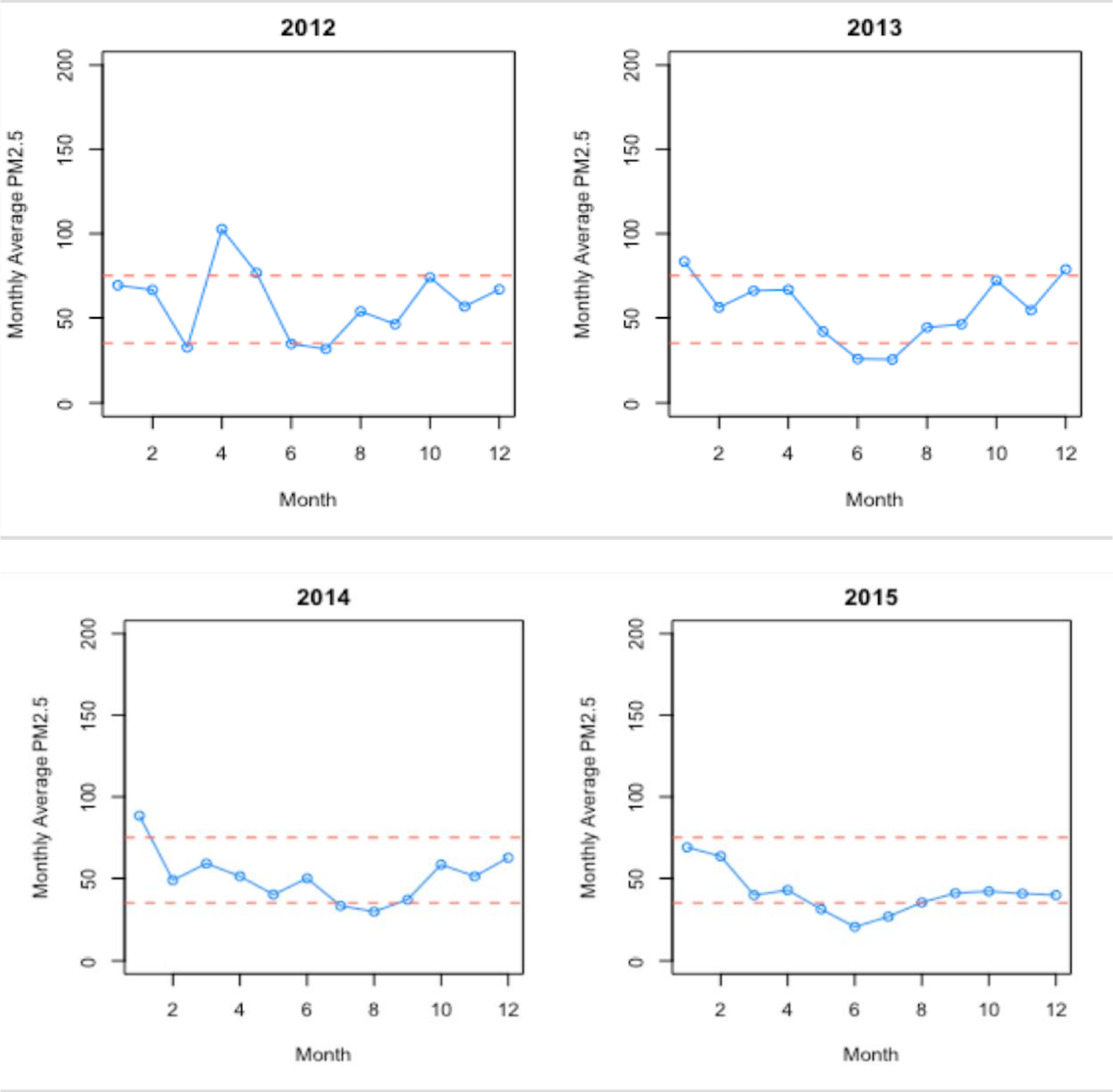


2. Monthly trend over years and within a single year

The monthly average PM2.5 plot below helps to better examine the yearly trend. Using PM2.5 as the single indicator of air pollution level, we can observe from Figure 2 below that the pollution is most severe during winter (December, January and February) and least severe during summer

(June, July, August, September) from year 2013 to year 2015. The average PM2.5 for most months were decreasing over time except for those of January and February. The air pollution level in those two months tends to be stable and our guess is that the traffic is heavier around the Spring Festival (an important for festival for Chinese to gather with families) when people need to travel to home.

Figure 2. Guangzhou Monthly Average PM2.5, 2012-2015



3. Guangzhou Air Pollution Trend

Overall, the amount of polluted and severely polluted days had been decreasing over time as illustrated in Figure 3.1. Alternatively, we observe from Figure 3.2 that the proportion of qualified days when people could go out without concerning on wearing masks had been increasing from 2012 to 2015.

Figure 3.1 Guangzhou Air Pollution Level, 2012-2015

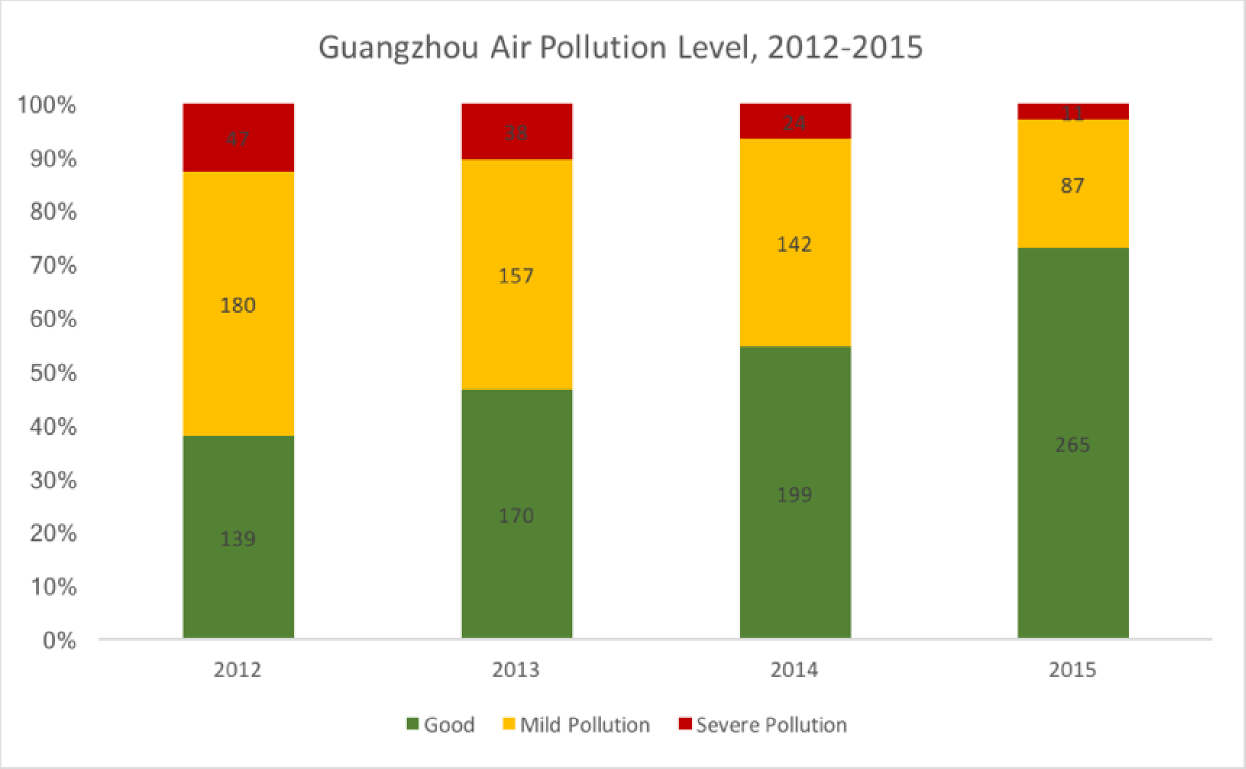
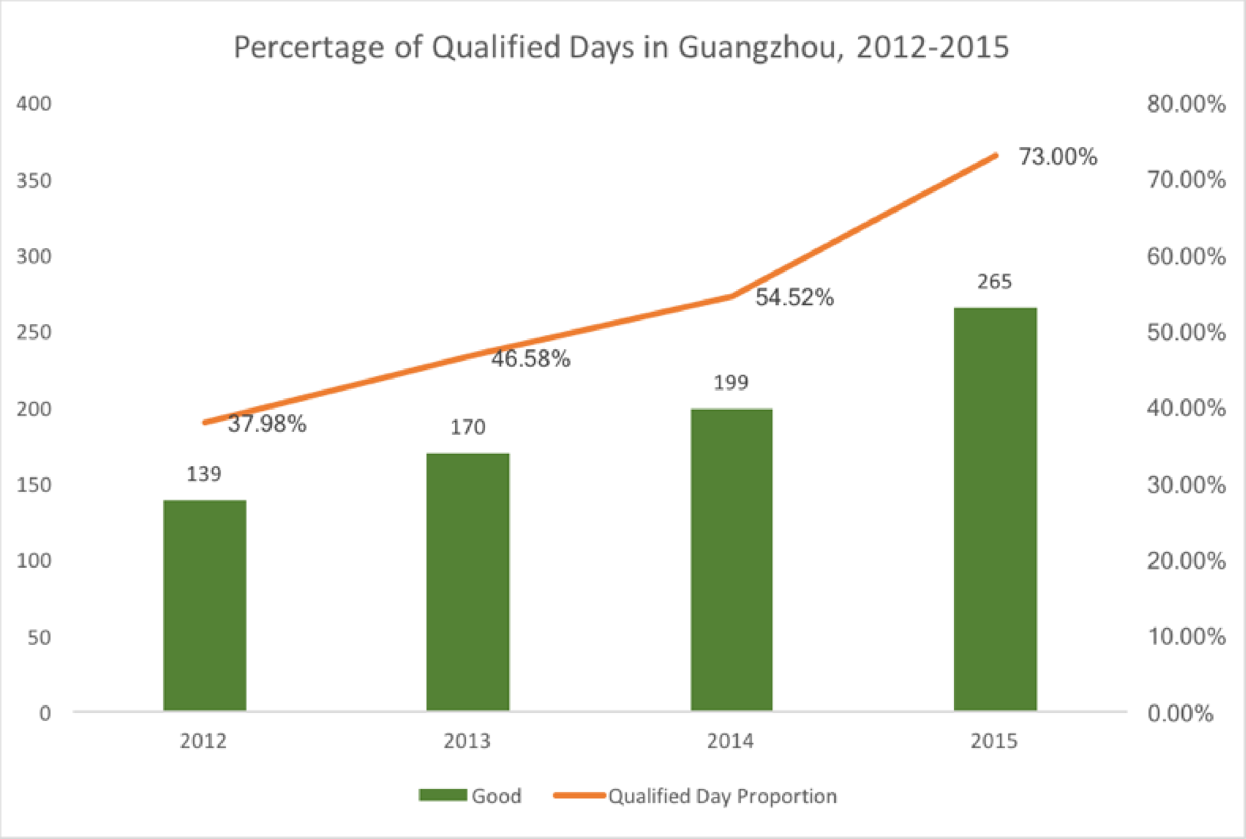


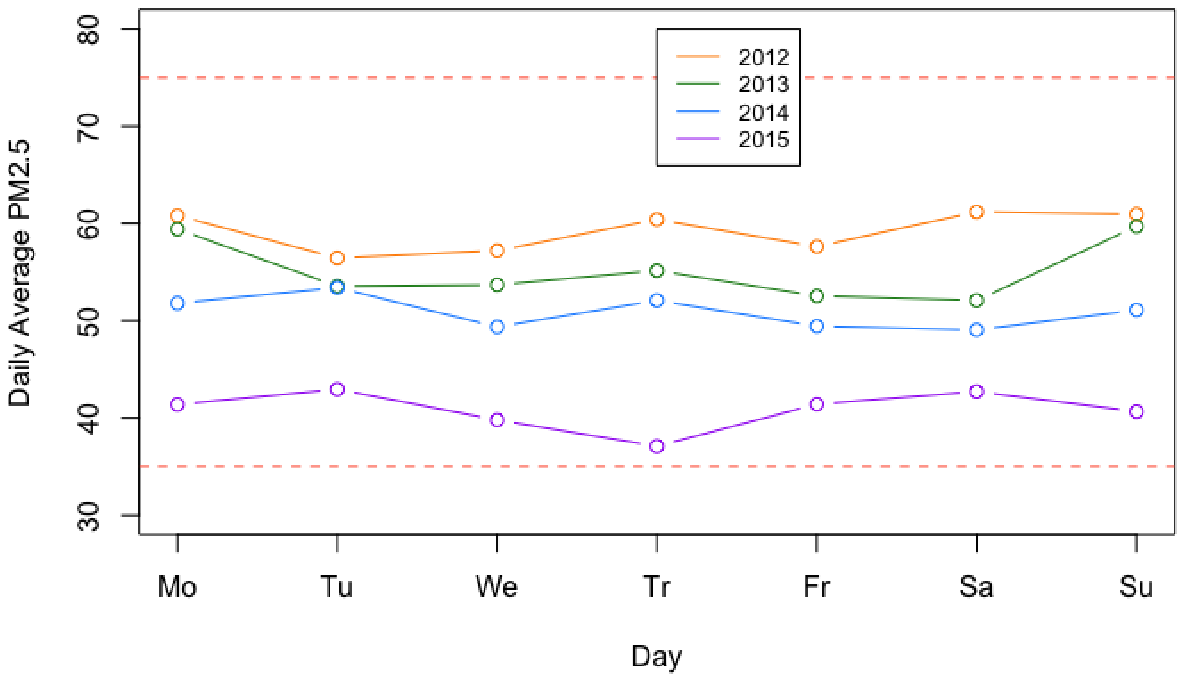
Figure 3.2 Qualified Days in Guangzhou, 2012-2015



4. Trend within a week

Since day is an important predictor for the PM2.5 level in Guangzhou, it might be helpful to examine the average PM2.5 level during weekdays and weekends. Considering automobile and industrial emissions the two primary resources for PM2.5 (Wang, Bi, Sheng, & Fu, 2006), there could be some difference on traffic and industrial operations in different days of a week. Although we expected weekdays to have slightly higher pollution than weekends since there is more traffic and industrial activities during the weekdays, the results suggest that there is no significant difference among different days in a week in terms of PM2.5 value.

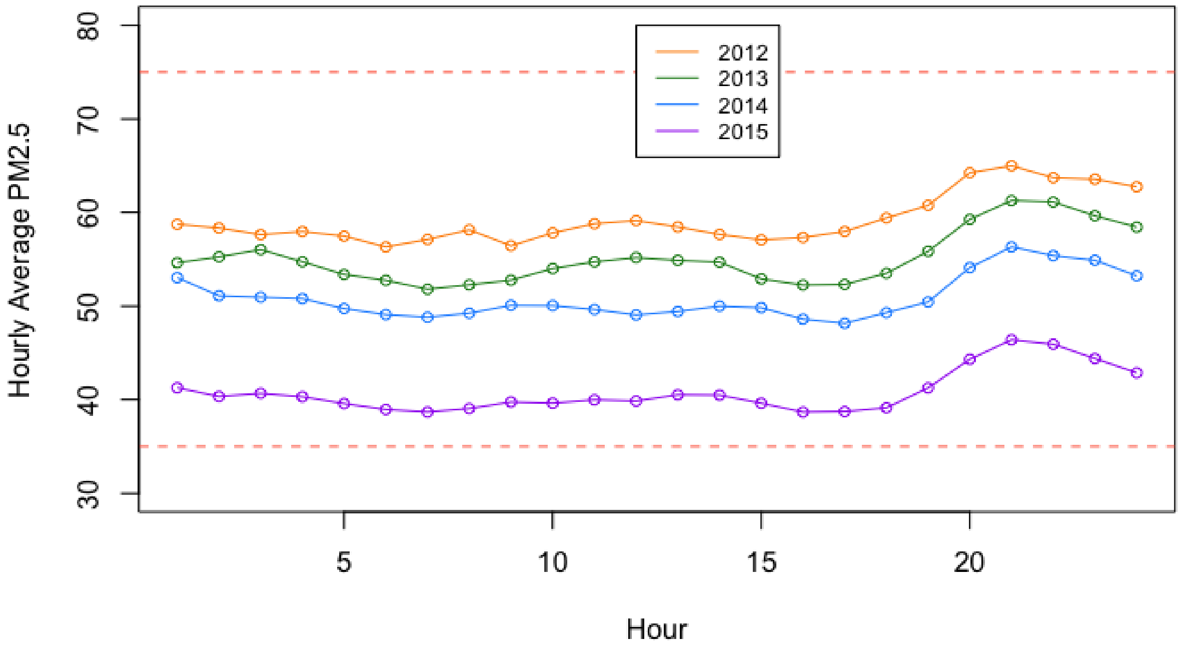
Figure 4. Guangzhou Air Pollution Weekly Trend, 2012-2015



5. Trend within a day

It might also be useful to examine during which period of a day the air pollution is worse than other time periods. Based on the data for four years, the air pollution tends to be observed more severe during early nights on average. Figure 5 indicates that the PM2.5 value was almost cosnat from the morning to the early evening. It started to increase in the evening and reached its peak around 8-9 pm and decreased overnight. Potential reasons for such pattern is that there is more traffic during evenings when people getting out from work and driving around for night activities.

Figure 5. Guangzhou Hourly Average PM2.5, 2012-2015



Conclusion

Based on our analysis, the amount of polluted days had been decreasing and the air quality of Guangzhou had been getting better from 2012 to 2015. It suggests that the policies on reducing air quality in the city were effective and we recommend residents in Guangzhou follow those regulations. We found that the air pollution is most severe during the winter of a year especially in January and February as well as during the early night of a day. We recommend residents to reduce travelling and outdoor activities during those periods or be prepared with masks when going out. The air quality is the best during the summer of a year and typically from

May to September. We think the public do not need to be too worried about the air quality during those periods unless they are going out during early nights.

Suggestion

Here we give some suggestions for the clients to conduct future research on the air pollution level of Guangzhou. Our first suggestions is: more research needs to be conducted on severely polluted days. One way to do additional research is that they can look at air pollution data in other more polluted northern Chinese cities, such as Beijing. If there are patterns for relationships between environmental variables and PM2.5 values during severely polluted days in Beijing, it’s likely that these variables may also result in high PM2.5 values in Guangzhou.

Our second suggestion has to do with our model having a small error rate but a high redundant rate too. We think that if our client can provide some other air quality measurement such as the value of SO2 and NO, we can try to improve our model with better result. Our third suggestion relates to the missing values in the dataset. We assumed these missing values happened randomly as we did not find any particular pattern of those values. However, it would be helpful if we could have additional information to verify that since our prediction could be more biased if the missing values did not occur randomly.

Reference

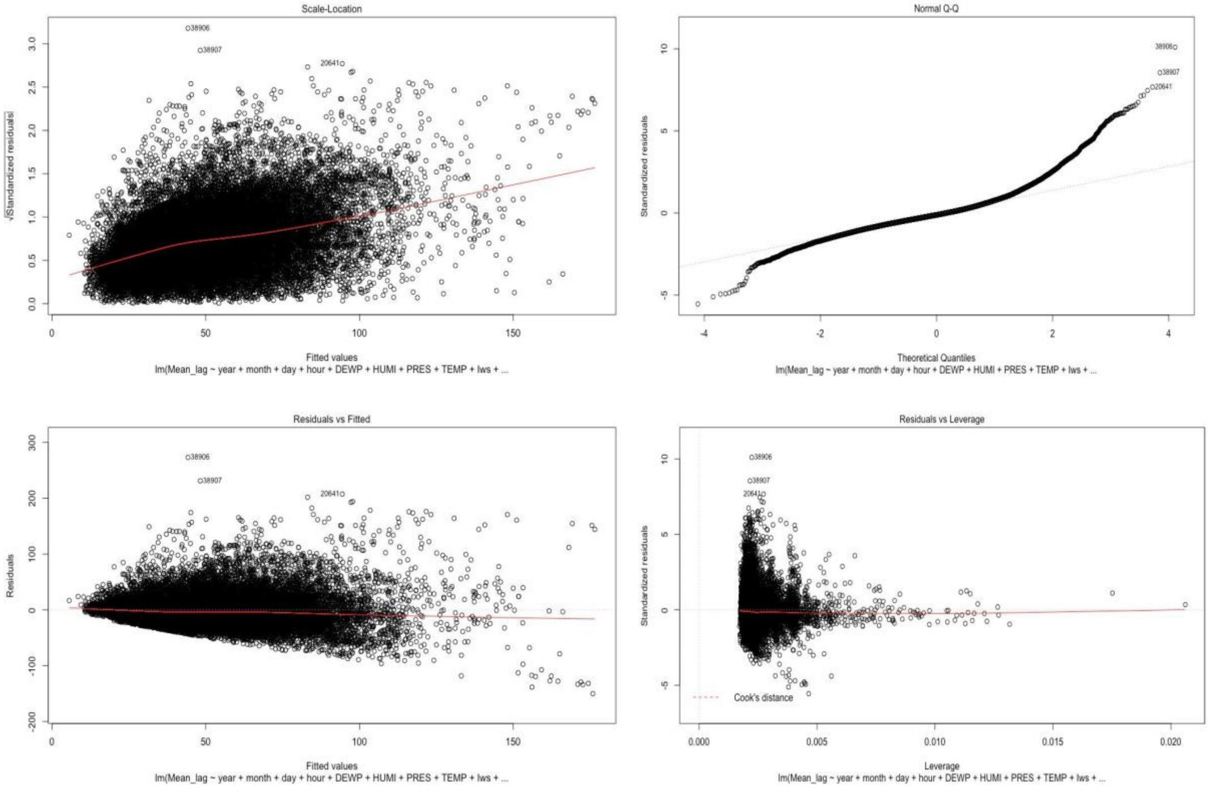
Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Wang, X., Bi, X., Sheng, G., & Fu, J. (2006). Chemical Composition and Sources of PM10 and PM2.5 Aerosols in Guangzhou, China. *Environmental Monitoring and Assessment; Dordrecht*, *119*(1–3), 425–39. https://doi.org/http://dx.doi.org/10.1007/s10661-005-9034-3

Appendence

Diagnostics

Below are the diagnostics plots for our chosen final prediction model. The QQ plot suggests violation of normality and the residual plot implies heteroscedasticity. Since we worked on a panel dataset which are usually more suitable for time series models, we are not surprised when the normality and constant variance assumptions of the linear regression model are violated. Additionally, we expect the previous days' PM2.5 values to be correlated with today's PM2.5 value and this suggests the independent observations assumption for the response variable to be violated. The cook's distance plot implies there are influential points in our dataset which are potentially the extremely high PM2.5 values. We did not remove them except for a few extremely large values because we need the data series to be continuous on discrete time intervals. Also, those influential points tend to be observations with large PM2.5 values that are in Group 3 (PM2.5 > 150) and account for only about 2% of the whole dataset. Removing these observations could highly affect our prediction of heavily polluted days (PM2.5 > 150). We also tried some transformation of our response variable in the regression model including Log transformation and Quadratic transformation. But they seem not to be helpful and we kept our original multiple linear model. However, because our goal of the model is prediction and classification, we believe it was acceptable to bypass these assumptions in our project. For future researchers, we suggest using rigorous time series models, such as multivariate time series model to capture the trend of PM2.5 values.


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
38906
38907
20641
0
50
100
150
Fitted values
lm(Mean_lag ~ year + month + day + hour + DEWP + HUMI + PRES + TEMP + Iws + ...
Normal Q-Q
Standardized residuals
-5
0
5
10
38906
38907
20641
-4
-2
0
2
4
Theoretical Quantiles
lm(Mean_lag ~ year + month + day + hour + DEWP + HUMI + PRES + TEMP + Iws + ...
Residuals vs Fitted
Residuals
-200
-100
0
100
200
300
38906
38907
20641
0
50
100
150
Fitted values
lm(Mean_lag ~ year + month + day + hour + DEWP + HUMI + PRES + TEMP + Iws + ...
Residuals vs Leverage
Standardized residuals
-5
0
5
10
38906
38907
20641
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
lm(Mean_lag ~ year + month + day + hour + DEWP + HUMI + PRES + TEMP + Iws + ...