The Report Committee for Nan Guo

Certifies that this is the approved version of the following report:

PM2.5 study:
Explore PM2.5 in Beijing using data mining methods and social media data

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor:

Byron Wallace

James Howison

PM2.5 study:

Explore PM2.5 in Beijing using data mining methods and social media data

by

Nan Guo, B.E.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2016

Abstract

PM2.5 study:

Explore PM2.5 in Beijing using data mining methods and social media data

Nan Guo, M.S.INFO.STDS.

The University of Texas at Austin, 2016

Supervisor: Byron Wallace

Air pollution is one of the worst outcomes from industrialization. Among other air pollutants, PM2.5 is believed to pose the greatest risks to human health as it can lodge deeply into people's lungs. This study focuses on exploring predicting aerial PM2.5 values from traditional pollutants and wind information using data mining and statistical models, including K-means, Markov chain, SVR, OLS models. Additionally, trending topics on social media is also considered to analyze how PM2.5 influences people's daily life. Considering Sina Weibo is the most popular social media in China, OLS and SVR models were also implemented with Weibo dataset. Predictions based on this study are expected to help government and concerned organizations do better in environmental protection.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 BACKGROUND

Air pollution has created one of the biggest environmental issues in China during the last several years. It has become a severe problem that is detrimental to human health and has been estimated to lead to more than 3 million deaths annually around the world by causing cardiorespiratory diseases such as lung cancer. According to China's ministry of environmental protection (MEP), the air pollution index (API) has covered six atmospheric pollutants, nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), PM10 (particles smaller than 10 um) and PM 2.5 (particles smaller than 2.5 um), since 2013. (State Environmeental Protection Administration of China)[1]. Among all of the pollutants, PM 2.5, which is called “Invisible Killer” by CCTV (China Central Television), stands out to have the most severe impact on human health.

PM, which is short for particulate matter, refers to the particles that can be suspended in the air for a long time. It includes solid particles such as dust and liquid droplets. Depending on their size, some large and dark ones are visible as dust and smoke, and other small ones can only be seen through a microscope. According to the US Environmental Protection Agency (EPA), particles less than 2.5 micrometers in diameter (PM2.5, which is shown in Fig. 1) are referred to as "fine" particles and are believed to

pose the greatest health risks. Because of their small size (approximately 1/30th the average width of a human hair), fine particles can lodge deeply into the lungs. (US Environmental Protection Agency)[2]. The sources that emit PM 2.5 include all combustion such as fossil fuel based power generation, vehicle engines, and domestic heating. As a result, the human-made sources of PM 2.5 weigh more than the natural ones. In addition to above direct emission, PM 2.5 could also be formulated from chemical reactions among gases such as Sulphur dioxide and nitrogen dioxides. It is true that when we breathe, the fine particles can reach the deepest regions of our lungs. Exposure to particles is linked to a variety of significant health problems ranging from aggravated asthma and one in this scale to pre-mature death in people with hard disease on the other. PM2.5 is worthy of the name "Invisible Killer."

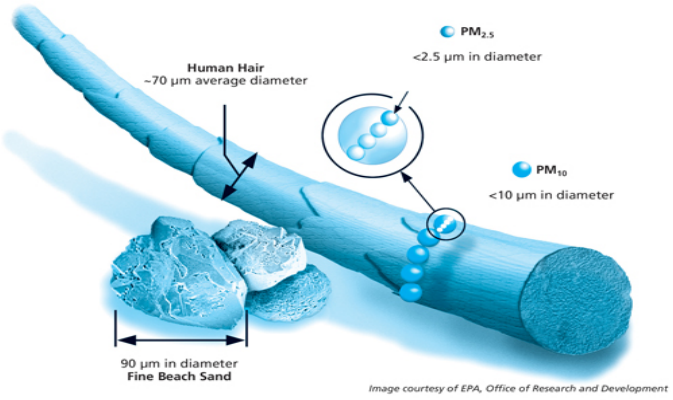

Figure 1: Illustration of PM2.5

Considering the fast pace of economic development and urbanization in China, the air pollution has drawn increasing amounts of attention from the public. For example, in 2013 Beijing had 58 days when the Air Quality Index (AQI) was higher than 200, which means “heavy pollution”. In December 2013 the east and central regions of China, which have more than 600 million people, experienced heavy pollution for more than two weeks. (Mei, Li, Fan, Zhu, & Dyer)[3]. Canadian researchers Aaron van Donkelaar and Randall Martin at Dalhousie University, Halifax, Nova Scotia, Canada, created a map (Fig. 2) by blending total-column aerosol amount measurements from two NASA satellite instruments with information about the vertical distribution of aerosols from a computer model. It is clear that most of areas in China are red, which is caused by industrial pollution. Areas in east China, where Beijing is located, are “super” red. Figure 3 shows a series of daily photos in Beijing in 2014. Most of the photos show that in that day the city was covered by haze. All of the evidence suggests that China is facing a big problem with air pollution.

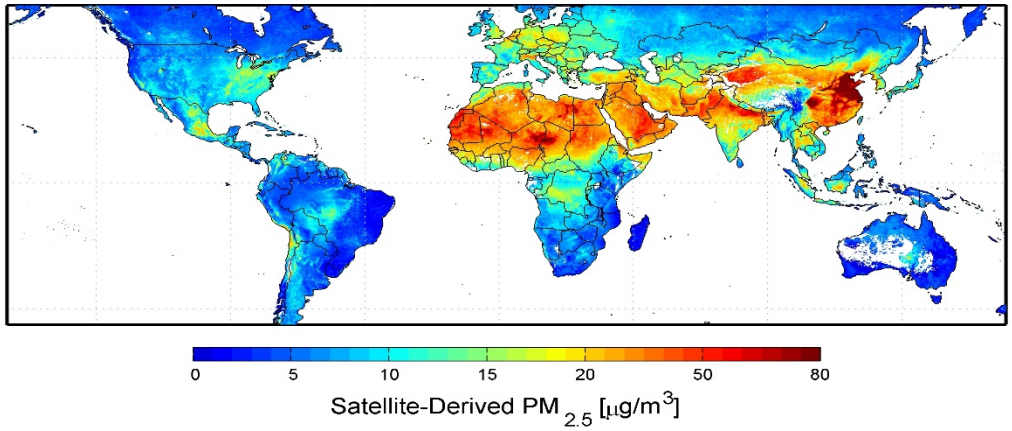

Figure 2: Global satellite-derived map of PM2.5 averaged over 2001-2006. Credit: Dalhousie University, Aaron van Donkelaar

Figure 3: Daily photos in Beijing, 2014

According to GUIDANCE ON PM2.5 MEASUREMENT UNDER DIRECTIVE 1999/30/EC, the mandate given by the Commission to CEN specified that the standardized PM2.5 measurement method "should be based on the gravimetric determination of the PM2.5 mass fraction of particulate matter collected on a filter under ambient conditions." Under gravimetric determinations the following methods are:

- weighing of filters before and after sampling, under well established procedures for weighing and filter conditioning
- tapered element oscillating micro-balance
- beta ray attenuation (not strictly gravimetric but assimilated as such)

Traditional weather forecasting is not enough anymore and people need more information than just temperature, humidity, and weather. People in China want more precise information about the air condition, especially PM2.5. As mentioned in Guidelines for Developing in Air Quality (Qzone and PM2.5) Forecasting Program, the air quality forecasting program can provide the public with air quality information with which they can make daily lifestyle decisions to protect their health. According to forecasting information, people will be able to make informed decisions regarding what transportation to take and where to exercise so as to limit their exposure to unhealthy levels of air quality. Organizations and companies can hold their planned activities on appropriate dates to attract more people to join. Now many big cities in China have

limited the use of cars according to their plate number, like people cannot drive cars with plate number ending in 1 and 2 on Monday. It is believed that the prediction of air pollution has the ability to help concerned parties control the pollutants, and also promote people's daily happiness.

Methods such as those based on optical methods (particle counting or nephelometry) are therefore not considered here. Since these methods cost money, a faster and cheaper way to estimate PM2.5 is desirable.

Already the ministry of environmental protection in China has set up air quality monitoring stations in more than 100 cities around China, and disclose an air quality index every day. Thus we can receive historical datasets of air condition and try to use them to predict the PM2.5.

Using social media data to predict the air pollution index is an inexpensive way to access air quality in different areas. As air pollution becomes one of the most important topics in China, there have been rich discussions among various social media platforms as social media gained popularity in China in last ten years. People are becoming more informed about environmental topics, and have a strong willingness to express their feelings through social media platforms. At the moment, Sina Weibo, which is similar to Twitter,

is the most popular social media platform in China, with more than 100 million messages posted daily. Furthermore, Sina Weibo is widely used across China and covers many different regions. All of the above features make Sina Weibo an ideal site to source related data. A previous work (Shiliang Wang, Michael J. Paul &Mark Dredze)[4] based on Weibo, has shown the high correlation between PM2.5 and some relative words in comments, like pollution, and breathe. Their work based on comments from the whole Weibo data made me think that whether I can build up models to study PM2.5 of Beijing from Weibo too. Therefore, as part of this project I wanted to explore how the PM2.5 of Beijing influences topics in Weibo and whether people's comments predict PM2.5 levels.

1.2 SUMMARY OF WORK

My prediction is based on two main datasets: (1) the historical dataset concerning air pollution measurements as well as wind factors (including SO2, NO, wind speed and wind direction degrees), and, (2) comments in the same period of time from Weibo. I wanted to explore how pollutant and wind data influence PM2.5, and how PM2.5 influence social media data. Using these datasets, I have created a series of methods to predict PM2.5 with relative factors and also forecast PM2.5 in the future.

For historical datasets, there are time series data of carbon monoxide (CO), ground level ozone (O_3), nitrogen oxides (NO), sulfur dioxide (SO_2), and particulate matter (PM2.5), which are identified by the original US Clean Air Act as five of the most common and

most important air pollutants (lead is also included in the most common air pollutants, however, Chinese organizations do not take it as seriously as other pollutants). Wind speed and wind direction degrees also caught my attention because common sense tells us that when wind speed is high, small pollutants will be blown away, as seen in the depiction of Tangshan, a large coal-mining site located in southeast Beijing (Fig. 4). Therefore I also include these two features into my analysis.

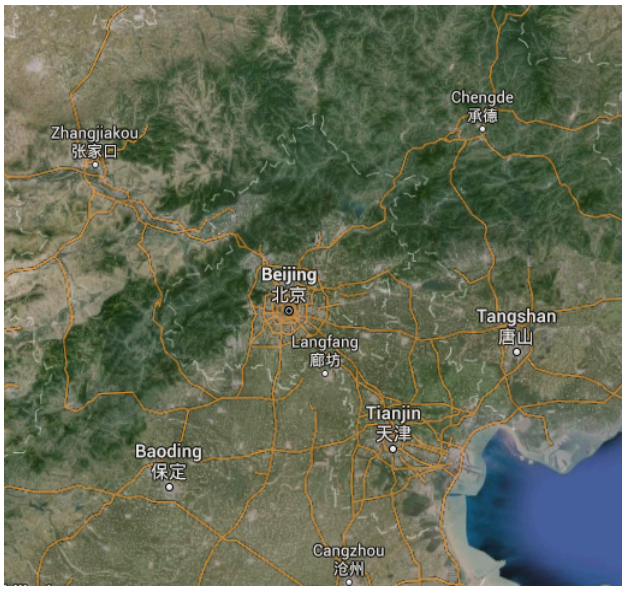

Figure 4: The location of Beijing and its surrounding

I analyzed and predicted PM2.5 with pollutants and wind data in two ways. The first is exploring how kinds of factors influence PM2.5. For this part of my analysis, I used Ordinary Least Squares (OLS) and Support Vector Machine (SVM) for regression to fit

and predict PM2.5 dataset. The other feature uses time series data of PM2.5 itself. I used the K-means model to smooth the data first and then used a Markov chain model to complete the prediction. After classification by K-means, continuous data got discrete into 4 levels. Each level can be taken as a status for Markov chain to predict the following status.

For Weibo data, I crawled comments and relevant information, e.g., the number of "like" comments and forwards from 01/01/2014 to 02/29/2016 under specific limitations. Then I prepared a document including key words, which are related to the topic of PM2.5. With the document, I first marked each comment as "PM2.5 relevant" or "PM2.5 irrelevant", and also recorded how many times each key word was hit each day. Next, I used the OLS and SVR models to fit and predict the PM2.5 from these comments, and calculated the correlation between PM2.5 and these key words.

The research reported here has several important limitations. First, we could do further analysis with wind speed and wind direction degrees more efficiently if we received the detailed information about industrial factories. Second, setting the number of centroids as 4 in the K-means model was done empirically, and might not be the best choice. Further research might improve prediction performance. Third, it is difficult to crawl "good" data from Weibo. Since there are millions of comments published everyday in Weibo, if I

randomly crawled comments they will overall provide little relevant information. However, if I collected comments from individual famous users who talk about air conditions or the weather everyday, the hit number of key words will be very high. Even though I was able to download many comments from Weibo, the approach is not guaranteed to be accurate. Ideally, one would use all comments available on Weibo. However even in this situation we cannot guarantee whether users meant what they were commenting or if their comments were ironic. Lastly, I used and updated the keywords from previous work (Shiliang Wang, Michael J. Paul &Mark Dredze)[4], which can be updated in future work. Compared to their work, I focused on PM2.5 in Beijing and implemented more statistical methods to explore how some features change according to PM2.5, while their work were based on the whole Weibo comments in a specific period of time.

In the non-media study with pollutants and wind data, I found that common pollutants and wind factors have a strong correlation with PM2.5. I can get a rough PM2.5 by using the OLS and SVR models. However, predicting PM2.5 by itself is not easy, especially for long-term predictions. If we just predict day by day, the results are reasonably good.

In the social media study, although social media has many comments that correlate with PM2.5 and air pollution, it is hard to abstract useful data from it. There are a handful of

reasons that might explain this. First, the crawling procedure will influence the results. Second, the scale of the data I collected may not have been big enough. Third, there may be an 'information delay' on social media.

2. Materials and methods

2.1 DATASET

2.1.1 Pollutant and wind dataset

According to MEP and the original US Clean Air Act SO2, O3, NO, CO and PM2.5 are five of most common pollutants. Due to the location of Beijing, wind speed and wind direction degrees were also taken into consideration. In order to make good use of wind direction degrees, the data was transformed into 12 ranks. I downloaded this data from aqistudy.com and wunderground.com, and post-processed this data.

2.1.2 Weibo dataset

In the social media study, to find relevant data from the massive number of comments in Weibo, I input “Beijing air” and “Beijing weather” into the filter to find hundreds of pages of comments. Then I stored all of the comments and meta-data, including the number of times each had been forwarded, liked and replied to. This information was collated into a CSV file.

2.2 DESCRIPTION

2.2.1 Pollutant and wind dataset

We can see the overview of PM2.5 first in Fig.5.

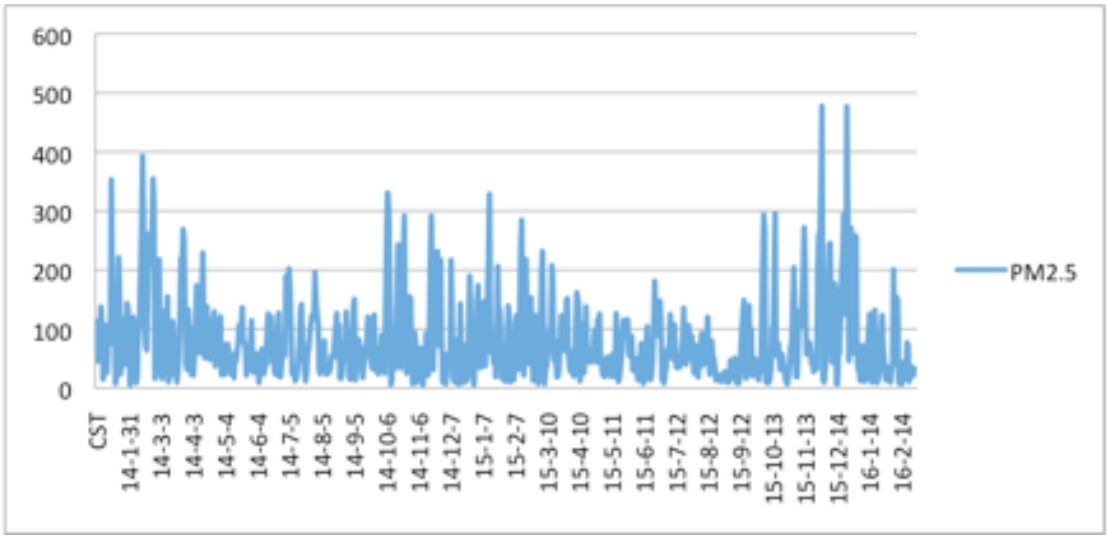

Figure 5: PM2.5 from 1/1/2014 to 2/29/2016 PM2.5 ranges from 5.2 to 477.5, and does not follow an obvious pattern.

Histogram figures are shown as below (Fig.6). The x axes shows the values of factor and y axes shows the amount. Some of the factors have outliers, which may influence some models during predicting.

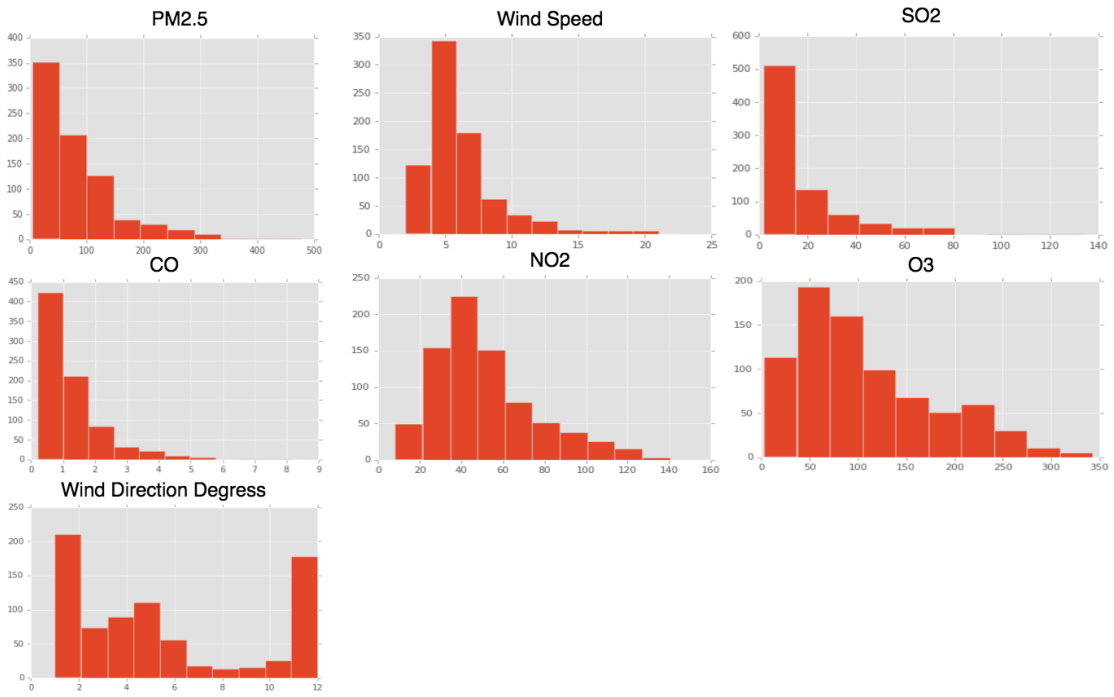

Figure 6: Histograms of PM2.5, wind speed, SO2, CO, NO2 and O3, showing the distribution of each dataset.

Basic information of all the data is shown in Table 1.

Table 1: Exploratory analysis of pollutants factors

	WindSpeed	WindDirDegrees	SO2	CO	NO2	O3	PM2.5	WindDir
count	790	790	790	790	790	790	790	790
mean	5.87	155.05	16.76	1.28	51.23	107.56	80.33	5.64
std	3.09	115.53	19.17	1.00	24.37	71.48	70.07	3.83
min	2	1	2	0.22	8.1	3	5.2	1
25%	4	54	4.125	0.63	34.125	56	29.2	2
50%	5	128	9.5	0.975	45.9	87	59.85	5
75%	7	281	20.85	1.5675	61.975	152.75	108.7	10
max	21	360	133.1	8.14	140.2	343	477.5	12

It is clear that the distribution of some of the data is skewed; we should thus be careful to choose and apply predictive models that are robust to outliers.

2.2.2 Weibo dataset

In Weibo I searched for two topics, namely, “air Beijing” and “weather Beijing” from January, 1st, 2014 to February 29th, 2016. Under that specific filter I retrieved 195,177 records. Table 2 shows an example record.

Table 2: A row of datasets crawled from Weibo

date	content	like	forwarding	comment
01-1-14	北京这周的天气是要干啥啊！春天来了啊！好热啊！@只是一个痴汉 你一走就要回归雾霾了……_(:3」∠)_ What's wrong with the weather in Beijing! Spring is coming! So hot! @somebody It's back to haze when you leave……_(:3」∠)_	0	0	5

The number of comments from each day is not the same I chose the smallest number of comments in a day as the standard, which is 65, to create a fair analysis. As a result, I have gathered 65 comments each day for 790 days in total.

2.3 Brief Introduction to Methods

The models I used for this research are the Ordinary Least Squares (OLS), Support Vector Machine for regression (SVR), K-means clustering, and Markov chain.

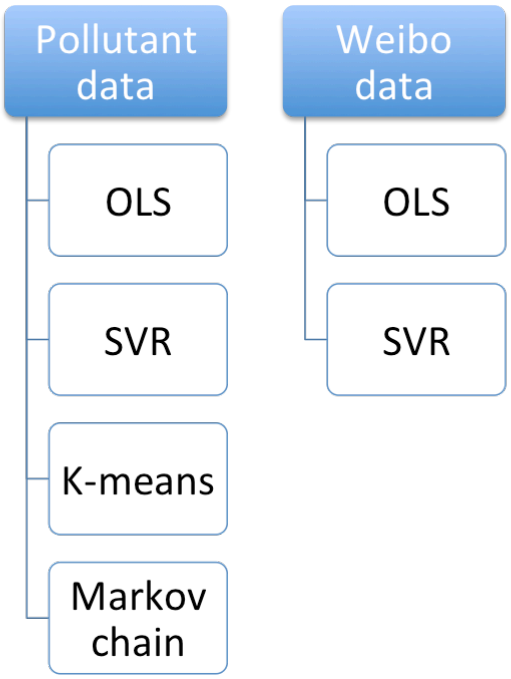

Figure 7: Models used in PM2. 5 study

2.3.1 Ordinary Least Squares

In statistics, ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in a labeled dataset (for which responses are known) and the responses predicted by the linear model.

2.3.2 Support Vector Machine for regression

Support vector machines are supervised learning models with associated learning algorithms used for classification and regression analysis.

A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola. This method is called Support Vector Regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

2.3.3 K-means

The K-means algorithm is an algorithm for putting N data points in an I-dimensional space into K clusters.

The data points will be denoted by $\{x^{(n)}\}$ where the superscript n runs from 1 to the number of data points N. Each vector x has I components x_i. We will assume that the space that x lives in is a real space and that we have a metric that defines distances between points. [6]

2.3.4 Markov chain

A Markov chain, named after Andrey Markov, is a stochastic process that undergoes transitions from one state to another on a state space. A Markov chain produces a sequence of random variables X_1, X_2, X_3, ... with the Markov property, i.e., that the probability of moving to next state depends only on the *m* preceding states and not on the previous states

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$, if both conditional probabilities are well defined, i.e. if $\Pr(X_1 = x_1, ..., X_n = x_n) > 0$.

The possible values of X_i form a countable set *S* called the state space of the chain. [7][8]

3. Data processing and model implementation

3.1 MODELS ON POLLUTANT AND WIND DATA

After collecting data, including the daily values of PM2.5, SO2, NO2, O3, CO, wind speed and wind direction degrees, I explored the relationship between PM2.5 and other factors. For this I used simple multivariate regression. Using multiple regression, we can estimate a coefficient for each factor, and then predict the PM2.5 in the future using these estimates and new observed data.

I was also interested in understanding how PM2.5 changes in the future according to its time series data. The Markov chain can help me to complete this prediction.

3.1.1 Multivariate regression

3.1.1.1 OLS

As I mentioned in the model introduction the goal is to get the minimum of e_i^2=(Yi-Yi)^2. Outliers may have a large effect on parameter estimates in multiple regression. According to the histogram in the data description (Fig.5), PM2.5, SO2, and CO contain outliers, so I used a log of the values to replace the original values in OLS model. Log transform can

transform count data to a normal distribution to drop outliers and satisfy assumptions of OLS regression.

Additionally, the values of wind direction degrees cannot be easily included in the regression model because the data is not linear. The values range from 0-359, encoding direction in degrees, however, 355 is very close to 0, because both of them point at north. Therefore I discretized all of the values of wind direction degrees to 12 classes according to their values. For example, 0-29 are transformed to class 1, 30-59 are transformed to class 2, and so on. This value is thus now a categorical variable. The model is seen below:

log(PM2.5)~log(SO2) + log(CO) + O3 + NO + WindSpeed + factor(WindDir)

Using 80% of the whole dataset as training data, after implementing the model, we can get the results as below(Table 3, and Table 4).

Table 3: The overall results of OLS for pollutant data

Dep. Variable:	log(PM)	R-squared:	0.805
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	158.9

Table 3 continued.

Date:	Sat, 30 Apr 2016	Prob (F-statistic):	4.25e-206
Time:	15:20:08	Log-Likelihood:	-282.40
No. Observations:	632	AIC:	598.8
Df Residuals:	615	BIC:	674.4
Df Model:	16		
Covariance Type:	nonrobust		

Table 4: The detail results of each variable

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	2.9861	0.108	27.722	0.000	2.775 3.198
C(WindDir)[T.2]	0.1515	0.060	2.538	0.011	0.034 0.269
C(WindDir)[T.3]	0.1715	0.067	2.553	0.011	0.040 0.304
C(WindDir)[T.4]	0.2721	0.062	4.383	0.000	0.150 0.394
C(WindDir)[T.5]	0.3683	0.061	5.995	0.000	0.248 0.489
C(WindDir)[T.6]	0.4663	0.074	6.275	0.000	0.320 0.612
C(WindDir)[T.7]	0.3404	0.105	3.245	0.001	0.134 0.546

Table 4 continued.

C(WindDir)[T.8]	0.2241	0.128	1.745	0.081	-0.028 0.476
C(WindDir)[T.9]	0.0666	0.115	0.581	0.561	-0.159 0.292
C(WindDir)[T.10]	-0.0594	0.097	-0.615	0.539	-0.249 0.130
C(WindDir)[T.11]	0.0057	0.074	0.077	0.939	-0.140 0.151
C(WindDir)[T.12]	0.0928	0.063	1.469	0.142	-0.031 0.217
WindSpeed	0.0182	0.009	2.089	0.037	0.001 0.035
log(SO2)	-0.0716	0.025	-2.850	0.005	-0.121 -0.022
log(CO)	0.9305	0.048	19.539	0.000	0.837 1.024
NO2	0.0123	0.001	9.191	0.000	0.010 0.015
O3	0.0027	0.000	9.441	0.000	0.002 0.003

C(WindDir)[T.X] (X can be any number from 2 to 12), is a categorical variable coded as 0 or 1, a one unit difference represents switching from one category to the other. R-squared is a statistical measure of how close the data are to the fitted regression line. We can see that here R-squared is 0.805, which is good. Then I used the model to predict the next 20% days. The result is seen below (Fig. 8):

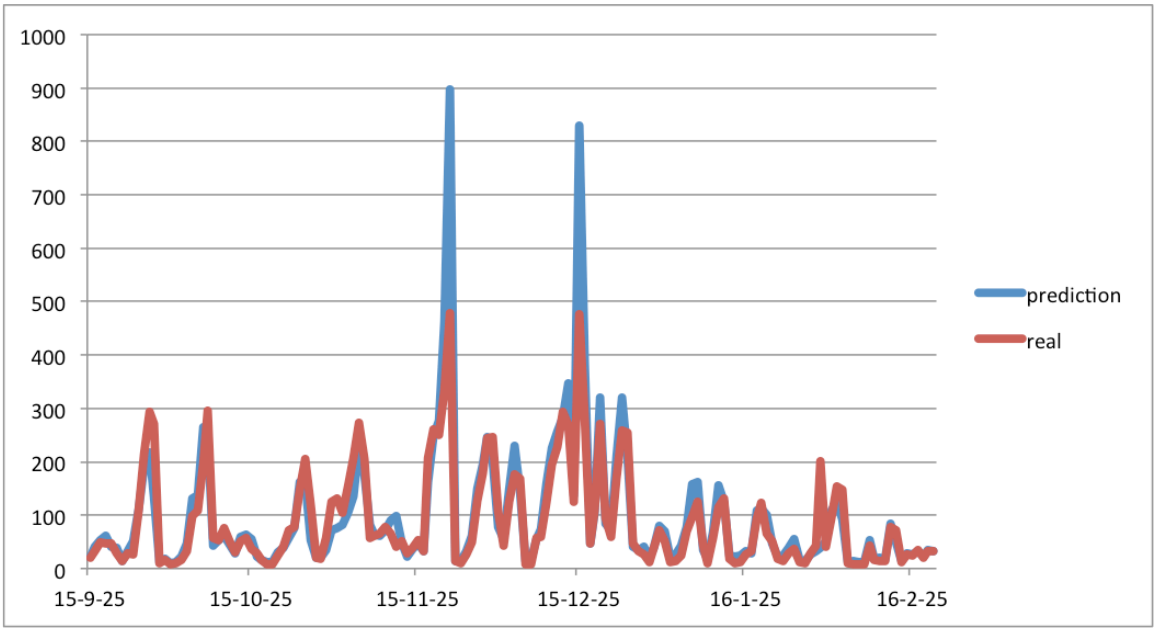

Figure 8: Predict PM2.5 in the next 158 days by OLS

I also calculated the mean absolute percentage error (MAPE), which is 28.7%. We can see the predicting model predict the trend well, but some of the predictions are overestimates and may influence the MAPE and other evaluation metrics.

3.1.1.2 SVR

In SVR, a non-linear function is learned via a linear learning machine applied to data mapped into a high dimensional, kernel induced feature space. SVR relies on defining the loss function that ignores errors, which are situated within a certain distance of the true value[9]. This type of function is often called an *epsilon intensive* loss function. Epsilon specifies the epsilon-tube within which no penalty is associated in the training loss

function with points predicted within a distance epsilon from the actual value. Therefore the value selected for epsilon will greatly influence the results.

I again used the same 80% of data as training data, and used the model to predict the remaining 20% of the days. I set epsilon as 0.01, 0.05, 0.1, 0.2, 0.5, and used the MAPE of the model as the decision criteria for selecting between these. However, the results will change, even when the same epsilon is used, because the approach is stochastic because of its cross-validation part. I ran the model 10 times with a different epsilon and found that when epsilon was 0.01 the MAPE was the smallest, and the variance of MAPE of each prediction is 0.02. Then, I continued running the model until the MAPE was lower than the average values. The MAPE I finally got from SVR was 31.6%. Figure 9 shows the predicted and actual values.

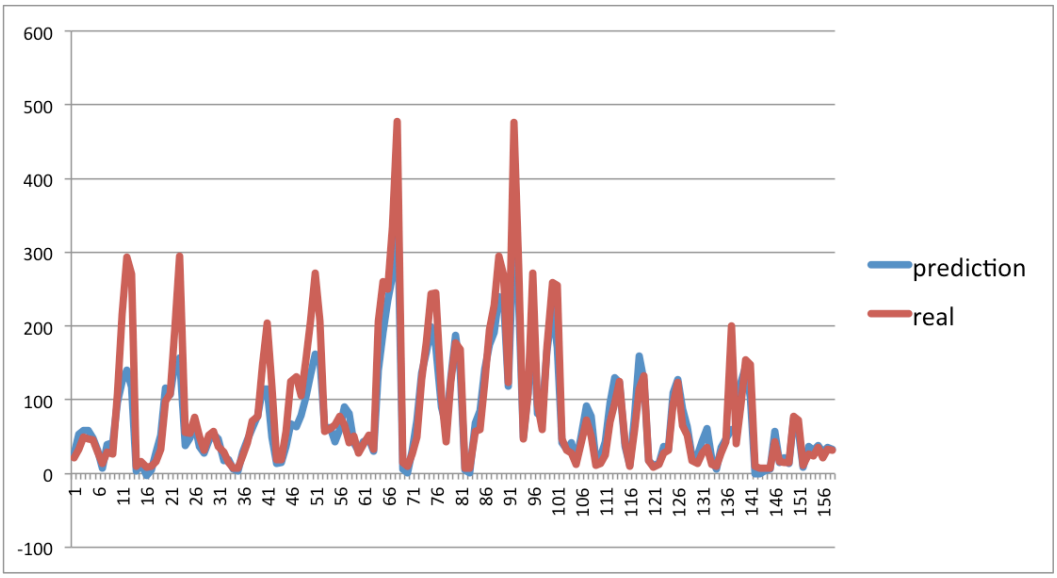

Figure 9: Predicted PM2.5 in the next 158 days using SVR

We can see that OLS had a better accuracy of the prediction than SVR, however, the prediction of SVR did not have the same outliers as the prediction in OLS did.

3.1.2 Time series predicting

Regression models tell use which predictors correlate with PM2.5, but we cannot know what PM2.5 will be in the following days from when we apply the model. If we could predict PM2.5 in the future independently, it would be ideal.

3.1.2.1 K-means and Markov chain

In most cases, the real values of PM2.5 have no intuitive meaning to people. For example, people cannot tell the difference between 300 and 500 in PM2.5. Most people

just need the level of the index to tell them whether the pollution is serious or mild and what the level is. I capitalize on this intuition and discretize the PM2.5 space.

To this end I use K-means. K-means is a clustering method to assign observations into k clusters, where observations are presumed to belong to the cluster with the nearest mean. So K-means can always produce tight clusters. However, fitting this model is stochastic: the initial centroids will influence the final centroids. Because I do not know what the best choice of initial centroids is I randomly set 4 initial centroids, ran the model ten times, sorted the values of centroids each time and calculated the average of the centroids in each position. The values in each same position did not vary a lot among ten implementations. Then I assigned all of the data into four clusters. The centroids, namely, 33.84, 100.84, 201.32, 332.95 are seen below (Fig.10):

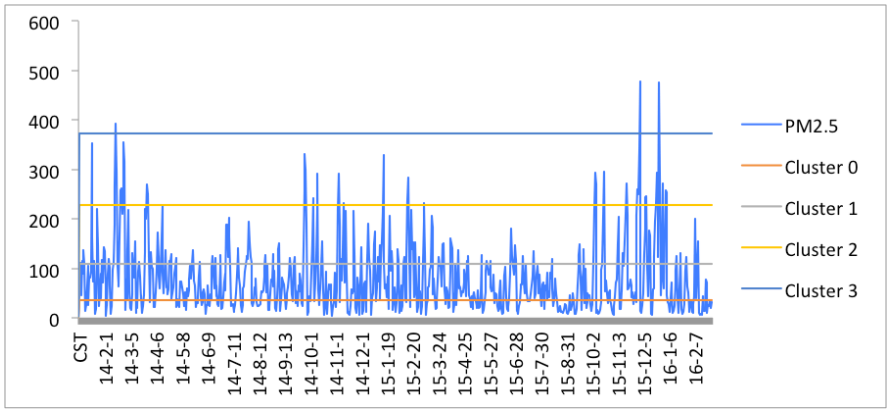

Figure 10: The values of mean of four clusters on original PM2.5 data

When all of the data was assigned to new clusters, I associated each cluster with different continuous level, from 0 to 3. The new dataset is shown in Fig.11. Compared to the original data, the transformed dataset kept the trend and became cleaner and simpler.

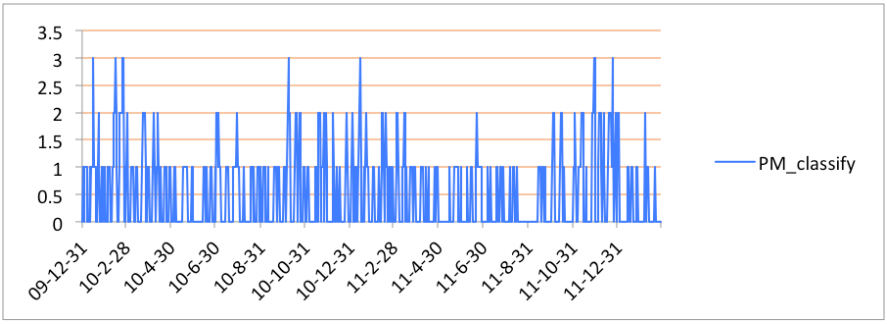

Figure 11: Transformed PM2.5 by k-means

In the next step, I implemented a first-order Markov chain to make predictions over time. I used the first 80% of the data calculated in the transition matrix:

$$R = \begin{bmatrix} 0.73 & 0.25 & 0.02 & 0 \\ 0.39 & 0.46 & 0.14 & 0.00 \\ 0.24 & 0.36 & 0.24 & 0.16 \\ 0.11 & 0.21 & 0.26 & 0.42 \end{bmatrix}$$

It represents the probability that one state will change into another state. The distribution over states can be written as a stochastic row vector x with the relation $x^{(n+1)} = x^{(n)}R$. So, if at time n the system is in state $x^{(n)}$, then k time periods later, at time n+k the distribution is:

$x^{n+k} = x^n R^k$

The start state is S0. Next I drew a chart of probability of the following steps (Fig.12). We can see that if we simply predict the following states with Markov chain, all of the predictions will be the same S0.

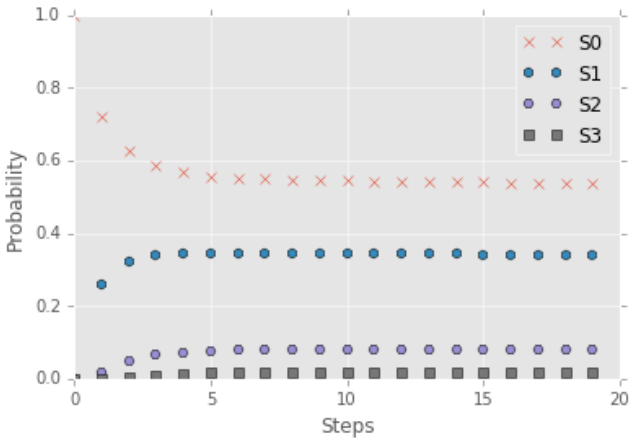

Figure 12: The probability of states in the following steps

To improve the accuracy of the prediction, I tried to predict steps day by day to see whether the result can be better. Because the transition matrix was calculated using 632 days, we can assume that the transition matrix will remain the same in a short period of time. I used

I used each of the previous day as the start state to predict the next state, and continued this 20 times. The predictions are seen below (Fig.13).

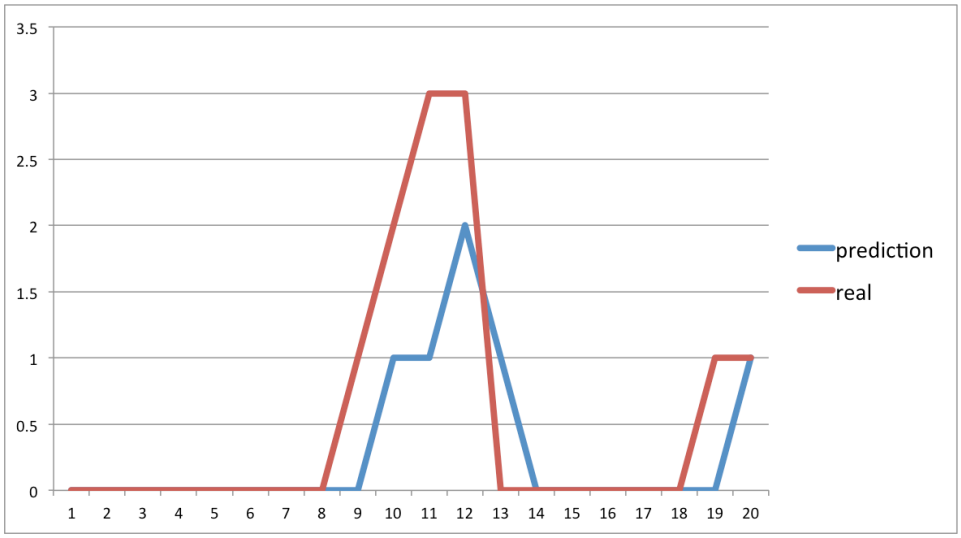

Figure 13: The day-by-day prediction by Markov chain

Although, the predictions have a similar trend as the real values, we can see there are some lags in predicting values. By using the Markov chain, we can get rough predictions, which can satisfy people's needs in some sense. However, it is hard to know where the inflection point is.

3.2 Models on Weibo data

3.2.1 Data handling

To standardize the comments, I took the minimum number of comments as a new filter and I retrieved 65 comments per day for 790 days, from 1/1/2014 to 2/29/2016.

First, I prepared a keywords list, including haze(“雾霾”), cough(“咳嗽”), grey(“灰”), PM2.5, pollution(“污染”), visibility(“能见度”), a word describes a vast expanse of whiteness(“茫茫”), throat(“嗓子”), sneeze(“喷嚏”) and breathe(“吸”). These keywords are based on Wang, Shiliang, Michael J. Paul, and Mark Dredze’s work[4]. They had “pollution”, “breathe”, “cough”. Additionally, I also made some updates according to my personal thoughts.

Second, I wrote code to automatically process the content. If any keyword was found in the content, I set the label of “mark” to “1”. I also recorded how many times each keyword appeared each day. Next, using the values of mark columns as filter, I calculated the sum of mark, forwarding, comments, and likes in each day.

Thus far, I had two new datasets, one including features; forwarding, mark, comment, likes and another including information pertaining to each keyword. For the first dataset, I used regression models to explore the relationship between PM2.5 and the metadata of comments. OLS and SVR models were implemented in this step. For the second dataset, I used the data to calculate the correlation table. I wanted to know how these keywords correlated with PM2.5.

3.2.2 OLS

To build the OLS model, I needed to know the distribution of each variable, because the model is sensitive to outliers. I drew the histograms for pollutant data (Fig. 14).

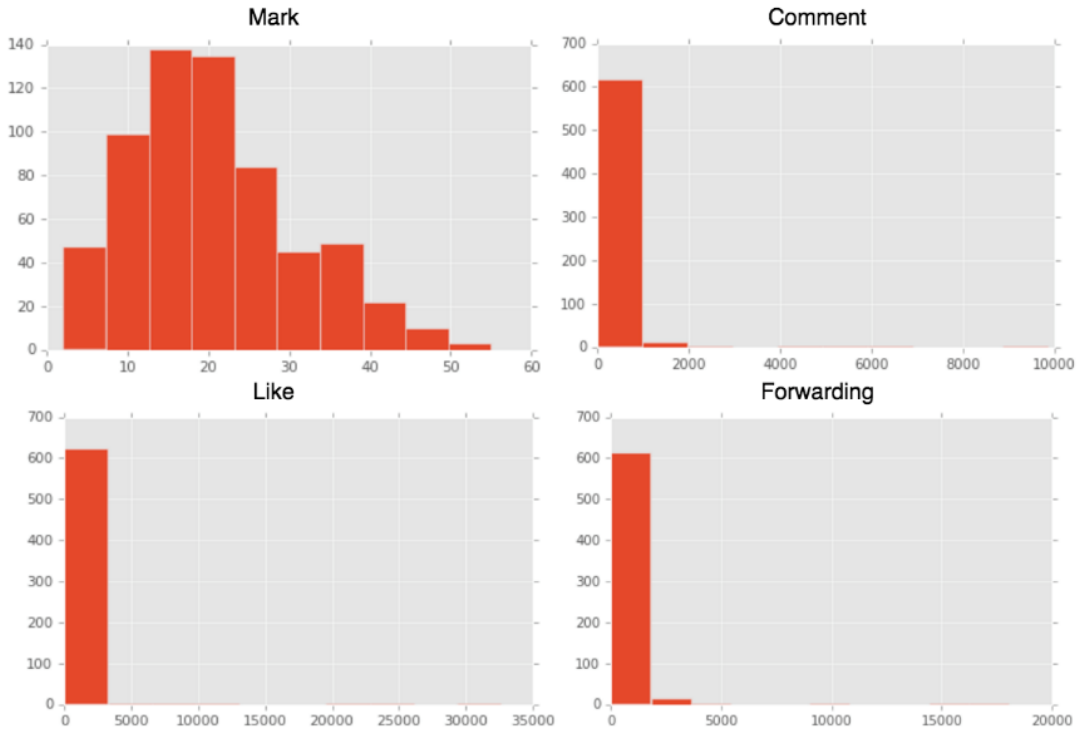

Figure 14: Histograms of mark, comment, like and forwarding

It became obvious that there were outliers in comment, like, and forwarding dataset. I used the log of values to replace the original one. My OLS model in Weibo is:

log(PM2.5) ~ mark + log(comment) + log(like) + log(forwarding)

Compared to the OLS in pollutant variables, I cannot simply implement the model because sometimes the number of comments, likes and forwardings might be 0, which is mathematically problematic because I am log-transforming. Therefore I added 1 to these three columns to remove 0s. The model results are below.

Table 5: The overall results of OLS for Weibo data

Dep. Variable:	log(PM)	R-squared:	0.022
Model:	OLS	Adj. R-squared:	0.016
Method:	Least Squares	F-statistic:	3.594
Date:	Sat, 30 Apr 2016	Prob (F-statistic):	0.00657
Time:	16:21:34	Log-Likelihood:	-792.18
No. Observations:	632	AIC:	1594.
Df Residuals:	627	BIC:	1617.
Df Model:	4		
Covariance Type:	nonrobust		

Table 6: The detail results of each variable

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	4.0600	0.096	42.320	0.000	3.872 4.248
mark	0.0121	0.005	2.658	0.008	0.003 0.021
log(comment)	-0.0084	0.052	-0.161	0.873	-0.111 0.094
log(like)	-0.0908	0.038	-2.365	0.018	-0.166 -0.015
log(forwarding)	0.0245	0.036	0.682	0.495	-0.046 0.095

The prediction results are also shown as below (Fig.15).

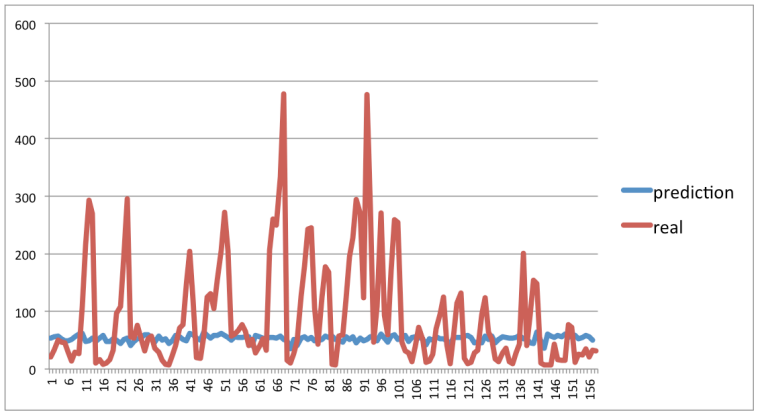

Figure 15: Results of OLS predictions for Weibo data

The MAPE is 1.22. From all the results; the table, the prediction and the MAPE, we can conclude that OLS cannot handle this regression problem. Another possibility is that the

data of like, forwarding, mark and comment have little correlation with PM2.5. In the next step, I analyzed further using the SVR model.

3.2.3 SVR

SVR model is non-parameter model, which focuses more on the physical distance. I wanted to see if the SVR model could be used to figure out the correlation between PM2.5 and Weibo content.

The predictions were occasionally smaller than 0, in which case I used log of the PM2.5 to replace the original values and re-fit the model. I cannot tell what combination of C and epsilon had the best results because the results of SVR changed in each iteration because of its cross-validation. I calculated the MAPE in each iteration, and when C, epsilon equals to (0.2, 0.01), the results were relatively good compared to other setting of parameters, although still poor. I ran the models for ten times, and the variance of each MAPE is 0.19. The results are below (Fig.16).

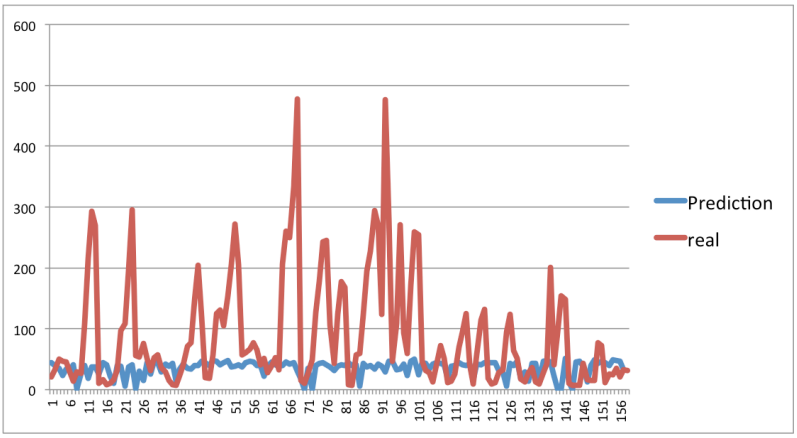

Figure 16: The prediction results of SVR using Weibo data

The MAPE result was 0.86, which was not good. Inspecting the results of OLS and SVR fit to the Weibo data, we can observe that features have little correlation with PM2.5 measures. With the data in hand it is not possible to make reliable predictions of PM2.5.

3.2.4 Correlation

For the keyword data, I calculated the correlation between PM2.5 values with each keyword. The correlation table is below.

Table 7: Correlation between PM2.5 and each keyword

	PM2.5_ words	Brea- the	Co ugh	Sne eze	Thr oat	Pollu tion	Gre y	Visi bilit y	a vast expan se of white ness	Ha ze	PM2 .5_ class
PM2.5_ words	1.00	0.21	0.1 0	0.05	0.0 9	0.41	0.07	0.0 3	0.05	0.1 8	0.08
Breathe	0.21	1.00	0.1 2	0.00	0.2 1	0.27	0.16	0.0 8	0.17	0.4 3	0.08
Cough	0.10	0.12	1.0 0	0.27	0.1 5	0.08	0.12	0.0 6	0.05	0.1 3	- 0.01
Sneeze	0.05	0.00	0.2 7	1.00	- 0.0 1	-0.03	0.04	0.0 5	-0.03	- 0.0 9	0.03
Throat	0.09	0.21	0.1 5	- 0.01	1.0 0	0.18	0.18	0.0 6	0.08	0.3 1	0.06
Pollutio n	0.41	0.27	0.0 8	- 0.03	0.1 8	1.00	0.17	0.2 5	0.14	0.5 0	0.08
Grey	0.07	0.16	0.1 2	0.04	0.1 8	0.17	1.00	0.1 0	0.07	0.2 4	0.06
Visibilit y	0.03	0.08	0.0 6	0.05	0.0 6	0.25	0.10	1.0 0	0.17	0.3 1	- 0.04
a vast expanse of whitene ss	0.05	0.17	0.0 5	- 0.03	0.0 8	0.14	0.07	0.1 7	1.00	0.2 3	0.02
Haze	0.18	0.43	0.1 3	- 0.09	0.3 1	0.50	0.24	0.3 1	0.23	1.0 0	- 0.01
PM2.5_ class	0.08	0.08	- 0.0 1	0.03	0.0 6	0.08	0.06	- 0.0 4	0.02	- 0.0 1	1.00

In the table we can see that the correlation between PM2.5 class (values transformed by k-means model) and each keyword is relatively low. We can see that certain words do have some correlation with PM2.5, including: breathe, haze and pollution. We can assume that on these days, people discussed the topic of PM2.5, complaining about hazing, breathing and the pollution problem. Further research into this should be performed.

In general, the Weibo data has had bad performances in regression models and in a correlation matrix. Regression models were expected to predict the PM2.5 by values of like, comment, forwarding and mark. The correlation matrix was expected to show the high correlation between PM2.5 and relative keywords. However, neither of them were successful. I believe there are various reasons for this and further discussion will be proposed in a later section.

3.3 EVALUATION

I implemented six data mining and statistical models, including OLS, SVR, K-means, Markov chainto fit and predict the PM2.5 with pollutant dataset and Weibo dataset.

For OLS models, I have reported the R-squared for each model. I also calculated the MAPE of the predicting values. Models performed well in the pollutant dataset. The R-squared is 0.805, which is high enough for predicting. The MAPE is 0.287, which is also

good, but still could stand to be improved in the future. Models in Weibo data performed poorly, and it could be said that OLS cannot be used for predicting PM2.5 with Weibo data. The R-squared is only 0.022, and the MAPE is 1.22. This is why I decided to use pollutant factors for predicting in the scenario analysis.

For SVR, I calculated the MAPE of the predictions. The performance of SVR with pollutant dataset is not bad, and the MAPE is 0.316. However, it is difficult interpret the SVR model, while we can easily understand how each factor influenced the results in OLS model. The performance of SVR with Weibo dataset is poor. The MAPE is 0.86 or more, and we cannot use it for predicting.

The K-means model suggested that there were 444 days in status 0, 251 in status 1, 72 in status 2 and 23 in status 3. Although I randomly ran the models ten times to come up with an average of centriods, the number of each status are not similar. Some outliers in PM2.5 may have caused this problem. However, classifying of PM2.5 is not only a mathematical problem, but also a social problem. Further research is warranted.

The Markov chain model was only implemented with pollutant dataset. I assumed the transition matrix would stay the same for the next 20 days. The accuracy of the predictions was 70%. If a one-level error is accepted, which means when the real level of

the PM2.5 is Status 1, the result that the predicting level is Status 0 can be regarded as right prediction, the accuracy is 95%.

4. Scenario analysis

Scenario analysis can present consciously several alternative future developments by considering alternative possible outcomes. In contrast to prognosis, scenario analysis does not use the extrapolation of the past. In this way, we can manually set values to factors, and see the possible results using our models. The OLS model for predicting PM2.5 with pollutants and wind information has had the best performance so far. It is also easy for multivariate regression models to control the variables, so I used it for scenario analysis. The main method here is control variable method. For instance, I often allowed the clear identification of cause and effect because only one factor is different at a time, meaning the effect of that single factor can be determined.

According to TABLE 1, we know the mean, 25% points, 50% points, 75% points, minimum and maximum values of all the factors. According to Fig.5, we can see the distribution of all factors.

Wind speed and wind direction degrees are nearly impossible for people to control, so I set several combination for their values, and other scenario analysis are based on the change of other pollutants. From the histograms of wind speed and wind direction degrees, we can see that in the most situations, the wind speed equals to five, and there are several peaks in wind direction degrees distribution, such as when wind direction

degrees equal to one, five and twelve. The one and twelve in wind direction degrees are similar, because both of them point at north. As a result, I explored other pollutants assuming the wind speed and wind direction degrees as (5, 1) and (5, 5). Additionally, I wanted to explore what the PM2.5 would be like if all the pollutants are in a low level, mean level, high level, and how each of them influence the PM2.5. I can get all of the basic information of pollutants from TABLE I. To explore how single factor influence the results, I set other factors as their values of 25% points, and the target factor as its maximum values. I made a dataset with these values, and implemented my OLS model to make a prediction.

First, I set wind speed and wind direction degrees as (5, 1). The results are as below (Table 8).

Table 8: Setting wind speed and wind direction degrees as (5, 1)

Scenario	Prediction
All mean	56.09
All min	5.62
All 25%	22.58
All 50%	40.14
All 75%	85.92
max SO2, other 25%	17.60

Table 8 continued.

max CO, other 25%	244.22
max NO2, other 25%	83.31
max O3, other 25%	49.03
All max	1525.81

It can be seenn that under the specific wind speed and wind direction degrees, in most cases, the values of PM2.5 are low. CO and O3 have a little effect on PM2.5, while CO has a significant positive correlation with PM2.5. SO2 seems to have a negative correlation with PM2.5 according to OLS model. No doubt that when all the pollutants are maximum, the PM2.5 is extremely high.

I changed the setting of wind speed and wind direction degrees to (5, 5). The results changed a lot (TABLE 9).

Table 9: Setting wind speed and wind direction degrees as (5, 5)

Scenario	Prediction
All mean	81.06
All min	8.12
All 25%	32.63

Table 9 continued.

All 50%	58.01
All 75%	124.18
max SO2, other 25%	25.45
max CO, other 25%	352.96
max NO2, other 25%	120.40
max O3, other 25%	70.87
All max	2205.13

We can see all the predictions get higher than they did previously. In this two-scenario analysis, all other factors are the same, except the wind direction degrees were changed. We can draw the conclusion that wind from degrees 5 may bring much more particulate matter. Degrees 5 stands for southeast of Beijing. We can look at Fig.3 again. As mentioned previously, Tangshan is a major coal-mining site and Tianjin has a lot of industrial factories. We have enough reason do further research about whether pollutants which are produced by these factories in Tangshan, and Tianjin are brought to Beijing with the southeast wind.

By scenario analysis, I have found that CO has the most powerful correlation with PM2.5 and that wind from southeast will cause major air pollution problems in Beijing.

Analysis is useful for government and relative organizations. To reduce the PM2.5 in Beijing, the government should track where the CO comes from and investigate whether most pollutants and particulate matter in Beijing are produced in southeast area, especially like Tangshan and Tianjin. Then it is possible to implement further environmental protection methods or publish laws to limit the emission of pollutants.

5. Discussion

Air pollution is one of the worst side effects of industrialization. Among other air pollutants, PM2.5 is believed to pose the greatest risks to human health as they can lodge deeply into our lung. This study focuses on exploring how traditional pollutants and other factors, including SO2, NO, and wind speed influence aerial PM2.5 level. Additionally, trending topics on social media can be used to analyze how PM2.5 influences people's daily life. Predictions based on this study are expected to help government and concerned parties do a better job in environmental protection.

Because PM2.5 is influenced by abundant direct and indirect factors, I did not have a list showing what factors determine the PM2.5 levels. In this study, one of my aims was to explore whether some of the most common pollutants have significant effect on PM2.5 mathematically. According to common and background knowledge, geography, wind speed and wind direction degrees may have effect on PM2.5 as well. Thus, they were taken into consideration.

In the first part of my project, relevant pollutants and other quantified factors including wind speed and wind direction degrees were regressed against PM2.5 using OLS, SVR methods. When PM2.5 data was put into time series, Markov chain was used for estimation. After implementing the K-means model, the accurate values of PM2.5 were

transformed into level, which were applicable for Markov chain model. When PM2.5 was predicted on a daily basis in Markov chain, the results had lag.

In today's society, social media is a major aspect in the daily life of many people. If we can find the correlation between the PM2.5 and social media, government and concerned organizations can make good use of the information, like posting useful suggestions for health protection on proper days, advertising for their products online and organizing outdoor or indoor activities. In this consideration, I tried to explore PM2.5 in social media, Weibo.

However, the approach to crawl data from Weibo was the first and one of the biggest problem I considered in PM2.5 research. Since comments talking about PM2.5 in Beijing were needed, I could not randomly crawl comments from Weibo. If I just crawled comments from some users who are famous in environmental protection, the results would be biased. If I randomly crawled data from users who are located in Beijing, the proportion of PM2.5 topic is extremely low. Finally, I used to search function in Weibo to search for "Beijing air" and "Beijing weather", which were indirectly correlative to PM2.5 topic. I downloaded all the comments in a specific period of time as the original Weibo dataset. After data standardization, the dataset was ready for further analysis.

In the Weibo dataset, posts where keywords appeared were marked, as well as the frequency that keywords appeared in comments everyday, and number of likes, forwards and comments of each record. OLS and SVR regressions were used to estimate the PM2.5, they also calculated the correlation. However, the results found were insignificant. A likely explanation is that although the sample included 65 comments per day for 790 days, it was not big enough. Also keywords list needed to be improved.

In the scenario analysis, I set reasonable values to pollutant factors, wind speed, and wind direction degrees, and tried to use control variable method to explore how single factor influence PM2.5. I found that compared to other pollutants, CO has the highest correlation with PM2.5. Additionally, north wind and southeast wind are the most common winds in Beijing. Compared to north wind, southeast wind will cause higher PM2.5. After looking at a map, Tangshan, which is a major coal mining site, and Tianjin, which is famous for its industrial factories are located in the southeast of Beijing. This evidence make us consider whether or not air pollution in Beijing can be controlled by limiting the emission from industrial factories in Tianjin and publishing more strict rules for coal mining companies.

6. Limitation and future work

For traditional pollutant and time series analysis, although the performances of OLS, SVR and Markov chain models are acceptable. These models, in addition to other models can have a better fit and predictions if further information is collected. However, there are four main limitations for pollutant analysis section. Some future work was proposed according to these limitations.

The first limitation was regarding the background research. Besides these most common pollutants, some other unpopular pollutants may have significant influence on PM2.5. More research should be done to find these potential factors.

The second limitation is to improve the performances of OLS and SVR models. I can transform original data into various formats, and set different combination of the parameters in the models. New models can be created for predicting. For example, extreme high values always appeared in OLS predictions, we can manually reduce the values under a specific condition.

Lastly, the setting the number of clusters used was four. Whether other number of clusters is better to use in real world situation should be explored further. Besides, there are kinds of clustering methods, like DBSCAN, or just setting quartiles as the centriods.

The basic idea for choosing K-means is that points in cluster have relatively short distance, and distance between clusters is relatively high. However, outliers also have serious influence on the results. As a result, further research should focus on the clustering methods.

For Weibo analysis, all the models have unexpected performance. There are two very difficult obstacles to be dealt with in processing.

The first obstacle is the actual data. It is nearly impossible to get completely useful data from Weibo. We cannot clearly define what useful data is in this case. The way we crawl data will cause bias. As I mentioned before, if I crawled comments from professional environmental protection users, there may be a lot of comments correlated to PM2.5. If I randomly crawled data from users who are located in Beijing, the proportion of PM2.5 topic is extremely low. Attempting to crawl data properly is a difficult task.

The second obstacle to overcome is the setting of the keywords list. There were two goals of keywords list. The first one is to see the correlation between PM2.5 and each keyword. From these results, we can see the popularity of the topic in some sense. The other one is to mark comments as PM2.5-correlated comment or PM2.5-uncorrelated comment. However, this is hard, because there may be semantic problems and lag. Language is complex. People sometimes use “strange” words to express their complaints on PM2.5,

while sometimes they used some highly correlated words in other situations. Additionally, people may post comments talking about the PM2.5 levels of the past or in the future. These problems need complex machine learning methods to detect and deal with.

References:

1. State Environmeental Protection Administration of China. (n.d.). *Weelky Report of Air Quality Index for 84 cities*. Retrieved from State Environmental Protection Administration of China:http://www.sepa.gov.cn/quality/
2. US Environmental Protection Agency. (n.d.). *Fine Particle (PM2.5) Designations*. Retrieved from EPA: https://www3.epa.gov/pmdesignations/faq.htm#0
3. Mei, S., Li, H., Fan, J., Zhu, X., & Dyer, C. R. (n.d.). *Inferring Air Pollution by Sniffing Social Media*. University of Wisconsin-Madison.
4. Wang, Shiliang, Michael J. Paul, and Mark Dredze. "Social Media as a Sensor of Air Quality and Public Response in China." *Journal of medical Internet research* 17.3 (2015).
5. Smola, Alex J.; Schölkopf, Bernhard (2004). "A tutorial on support vector regression" (PDF). Statistics and Computing 14 (3): 199–222.
6. MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.
7. Norris, James R. (1998). Markov chains. Cambridge University Press. Retrieved 2016-03-04.

8. Markov chain. (n.d.). Retrieved May 05, 2016, from https://en.wikipedia.org/wiki/Markov_chain#cite_note-2

9. Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." *Statistics and computing* 14.3 (2004): 199-222.