**Iteration 4**

Jing ([jluo669@aucklanduni.ac.nz](mailto:jluo669@aucklanduni.ac.nz))

GitHub: <http://github.com/jluo669/Iter4.git>

**Step1: Business Understanding**

* 1. **Business Objectives**

Although developing quickly, Beijing is still suffering from pollution and many citizens are living in a terrible environment (Zhang et al., 2005). They also mentioned that in these elements, pm2.5 is one of the best indexes, which could display the quantity of environment. In the research, we will use data mining to try to minimize the value of pm2.5. The purpose of our research is:

1. By finding the key factors that affect pm2.5, we could find out some right methods of calculate the level of pm2.5 and the error within 10%.
2. Help these people, who live in Beijing, have a good understanding about their living air quality.
3. As one of the most representative cities, Beijing could be a good example for those cities, which have the same problems. So, the result could also applicable for other cities.

Success criteria: firstly, we need to make sure that these elements we found are the one, which really raise the pm2.5 value. Next one is that the solution we found could reduce pm2.5.

* 1. **Assessing the Situation**

The reason I choose this topic is that environment problems are worried by many countries. So, in order to solve this question, I found a dataset from a website, called Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>) and it is could be used publicly. This is an open source website and do not need some extra payment. There are no security issues, legal issues and other issues from using. This is not a group work. I finish this report by myself. All the analysis of this dataset is based on my understanding and I did not anything for this dataset. I only download them from the website and do some analyze, which I learned from INFOSYS 722.

The dataset’s name is “Beijing Multi-site Air-Quality Data Data Set” and they are stored in the Excel. This dataset includes the Aoti station’s PM2.5 in Beijing per hour every day from 2013 to 2017 with 35065 rows. The reason I use it is that it is no doubt that Beijing is one of the modern cities all around the world. However, with the fast development, it faces a horrible pollution in this recent year. Their citizens have to wearing masks when they go out. So, it could be a good representativeness. Moreover, it includes many factors that influence the level of pm2.5, such as NO2, SO2, temperature, raining. From analyzing these factors, we could find out how much they would infect pm2.5 and which is the most important factors we need to be more care about.

Some risks I met during the project:

1. Using these new technologies, like AWS, Jupyter, SSH, GitHub. These technologies are quite new for me that I never used them before, so it is quite hard for me to understand them and use them into my project in that such short time.
2. Programming: because my background is not computer, programming is not my specialty. During the programming, understanding code and fixing bugs even almost killed me.
3. COVID-19: due to the sudden onset of epidemic disease, it also makes me have less communication with lectors, tutors and classmates. Many problems cannot be solved in time, which also brings me a lot of difficulties.

In this research, some can happen by accidents. For example, duo to the virus, we need to stay at home and take lectures online, while because of the network traffic, sometimes my internet cannot work. That forced me postpone my original plan. Otherwise, at the end, I didn’t get the results that I wish, which I will be much frustrated. So, the only thing I can do is to try my best and hope for a good result I will get.

**1.3 Determine Data Mining Objectives**

Under the purpose that has been mentioned at the beginning, the important of this project is that to conclude what influence factor would affect the value of PM2.5 from the dataset. From analysing the data, we want to achieve that:

1. Finding a perfect model to describe the relationship between PM2.5 and these influence factors.

2. Analysing how much are these factors infect the level of pm2.5? And based on some research, we could try to find some ways to reduce these factors one by one, which will also decrease pm2.5.

Influence factors includes PM10, SO2, SO, NO2, CO, O3, temperature, rain, wind speed. The target is the value of PM2.5

Since almost all data is recorded by number, the type of this dataset is regression. The goals are using python to test each element how much effect on the valve of pm2.5, and then we could get an accuracy number. If the influence factor is greater than 0.001(0.1%), we can say that this factor has an impact on PM2.5. If it goes well, we can reduce PM2.5 and do not affect the city’s development. After this paper, we could find the solution to reduce the value of PM2.5.

* 1. **Project Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Time** | **Resources** | **Risks** |
| Business understanding | 2 days | IBM handbook, dataset | Having a good know about the question and requirements. |
| Data understanding | 2 days | AWS, Jupyter, SSH, PySpark, Spark, IBM handbook, dataset | Hard to have a professional perspective on the dataset. |
| Data preparation | 2 days | AWS, Jupyter, SSH, PySpark, Spark, IBM handbook, dataset | Using and understanding programming |
| Data transformation | 1 days | AWS, Jupyter, SSH, PySpark, Spark, IBM handbook, dataset | Using and understanding python programming |
| Data-mining method selection | 1 days | Google Scholar, IBM handbook, dataset | Research, select the appreciate method |
| Data-mining algorithm | 2 days | Google Scholar, IBM handbook, dataset | Research, select the appreciate algorithm |
| Data mining | 1 days | AWS, Jupyter, SSH, PySpark, Spark, IBM handbook, dataset | Building models. |
| Interpretation | 2 days | AWS, Jupyter, SSH, PySpark, Spark, GitHub, IBM handbook, dataset | Analyze the results and models. The way to submit assignment. |

**Step2: Data understanding**

**2.1 Collecting Initial Data**

The dataset I use is an existing data in the web, called Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>), which is a public data. In the dataset, I think the most essential columns is pm10, and then is the content of CO, NO, O3. Perhaps the temperate, precipitation, pressure would have some affect, but there could be no relationship with the number, year, date and hour columns. However, like CO, NO, O3, these gases are collectively called toxic gases, so we could discuss them together. The data type in the dataset are continuous and categorical.

**2.2 Describing Data**

The dataset has 18 columns and 35,065 rows and store in the excel. The following picture shows the type and description of each column. In this dataset, the column of PM2.5 will be the target and others are the influencing factors. There is a special column, station. Because we just collect one station in Beijing, the station column will be showed the same value, “Aotizhongxin”. For the rainy column, If the data is 0, that means it was not rain. If there are some value in the block, it represents the precipitation at that moment.

|  |  |  |
| --- | --- | --- |
| **Columns** | **Types** | **Description** |
| No. | Integer | Row’s number |
| Year | Integer | The year of the row |
| Day | Integer | The day of the row |
| Month | Integer | Row’s month |
| Hour | Integer | Row’s hour |
| PM2.5 | Integer | PM2.5 concentration |
| PM10 | Integer | PM10 concentration |
| SO2 | Float | SO2 concentration |
| NO2 | Float | NO2 concentration |
| CO | Float | CO concentration |
| O3 | Float | O3 concentration |
| TEMP | Float | Temperature |
| PRES | Float | Atmospheric pressure |
| DEWP | Float | Dew point temperature |
| RAIN | Float | Rainfall and precipitation |
| WD | String | Wind direction |
| WSPM | Float | Wind speed |
| STATION | String | The name of the air-quality monitoring site |

**2.3 Exploring Data(可视化)**

From the dataset we found, we could do some hypotheses.

1. it is easily for us to know that the column of No., year, day, month have no relationship with pm2.5, so during the research, we could ignore them.
2. The pm10 would have a significate impact on pm2.5.
3. About these toxic gases, if they increase, the value of pm.2.5 will increase.
4. For the temperature, wind speed, raining, as we all know, they trend seems like these toxic gases, which would increase the level of pm2.5.

And also, from these codes, they could help us have a good view of our dataset,

A screenshot of a cell phone

Description automatically generated

* 1. **Verifying Data Quality**

Roughly looking at the dataset, because most of columns are numbers and there are not extremely large or small numbers for each column. We could say that our dataset doesn’t have some errors values.

However, it is obviously that there are some missing data in the dataset. As it shows of the followings. I think the reason why they missed the data is that

1. they forget to record them.



1. Simplify the recording step. Because each data shows an attribute in per hour, the changing would not be much great, compared with the above one. If the data shows “NA” maybe because they hold same value as the above one.

图片包含 游戏机

描述已自动生成

From calculating the NA values for each column, we could see that these missing data are a very small part, which only occupy less 5%. The results of missing data for these columns are 718, 935, 1023, 1776, 1719, 20, 20, 20, 20, 81, 14. Compared with the size of dataset, these figures are completely negligible.

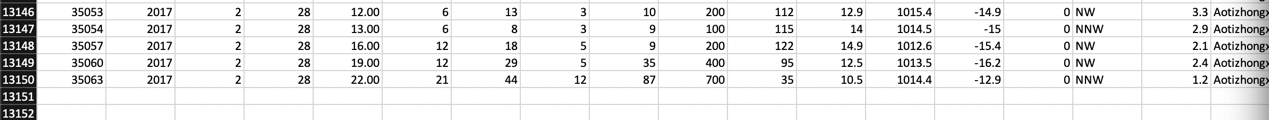
A screenshot of a cell phone

Description automatically generated

**Step3: Data Preparation**

**3.1 Selecting Data**

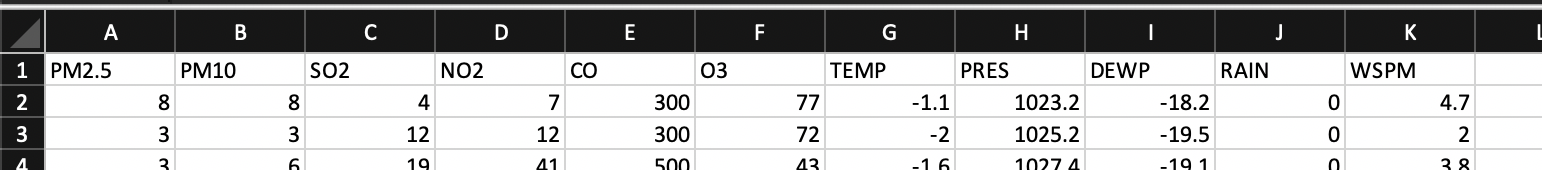
Because our dataset is so lager, it is necessary for us to reduce it size. It would increase the speed of our following steps. So, for my decision, I will only choice eight rows for each day. That means I will delate these rows, except for the hour equals to 1,4,7,10,13,16,19 and 22.



* 1. **Cleaning Data**

Our goal is to predict the level of pm2.5, so for some columns, it is clear that these infectors will not infect pm2.5 value, such as No, year, day, hour. And of course, for wd and station column, they do not contribute to our analyze. Because the dataset only shows one station in Beijing, all the data in station column are same. They are no help for our following step, so we will delete it. Another one is wind direction: maybe the wind level would impact pm2.5 somehow. However, in my opinion, their direction does not affect the results, but not sure, so we will keep it and do some future learning. If they are no relation with pm2.5. I would consider delete them.

So, for these useless columns, we will delate them from our dataset and do not consider them in the following steps.



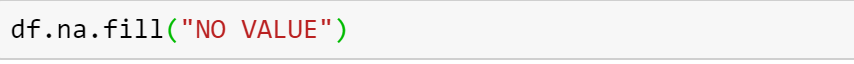
**3.3 Constructing New Data**

In the dataset, there are some “NA” recordings. We have already known that they represent the missing data. However, in the system, these are recorded as a string value, called NA, instead of null. So, after reading in the python, it could not read it as a null value. The system thought it is a value, called “NA”. That brings some difficulties for PySpark to recognize and analyze. Therefore, if we want to formatting data, we need to fix it firstly. After identified, all string “NA” and missing data are equals to null value. Although in our new excel, they still show NA, they could be read by python with the meaning of null, which will much convenience for our following steps.

Although after deleting, our dataset has only 13150 rows, there are still a huge data need to be storage. how

* 1. **Integrating Data**

In the rest dataset, there some missing value. For my preparation, I prefer to set them to “no value”, because as we just mentioned that, they only occupy a small part in each column, which we even could ignore them. However, if we consider them to 0, they would create some extra error during the analysis, so, we just ignore them and set them to none value. That would be the best choice.



* 1. **Formatting Data**

After the following step, we could put these new data in the new file, which would convenient for us to the following step, because if not, python will store them in the memory. However, our dataset is so big with more than 30,000 rows. That is very harmful for my laptop. Besides, I even do not have enough areas for storing them. We are going to build a new file to store our data.

**Step4: Data Transformation**

* 1. **Reduce the Data**

There are some other values could be delated in the dataset. The obvious one is none value in pm2.5 column. Our aim is to predict pm2.5 from these infectors and compared them with the recording to see the accuracy, thus to analysis our modeler is suitable or not. However, if there are no recording in the attributes, they are no sense for us to predict and analyze, so we will delete these columns.

* 1. **Project the Data**

For the rest of data, we all save them in the file, called “test”.



**Step5: Data-mining Method Selection**

**5.1 Match and Discuss the Objectives of Data Mining to Data Mining Methods**

There are many data mining method nowadays. For the followings, we will discuss some of the most common methods.

The first one is called classification. Ahmed and Elaraby said that classification usually use to predict the results labels (2014). In the report they said the labels could be “yes” and “no”, “safe” and “risk”, which are already been set up before the analyzing. For example, before the loan, the bank will evaluate the Gao of the loan to this person. There we could use classification method. Because we already have the result labels: “high” or “low” and the result must be one of these. They should have the value and the result do not have some extra value.

The second one is regression. it usually uses to predict the number, like temperature, profit, etc. This method needs a known dataset. Through this method, they could find the relationship with the result and predict the value. This predict value would compare with the input value and use variance to show the correction (Bailey, Muth &Nourse, 1963).

The last one is clustering method. This method put data into a group or clusters by using some algorithms. It is different from classification, because its label in unknow. They method would analyze the characteristics. This data mining method could be used to recognize the image and web search (Zhang, Ramakrishnan &Livny, 1996).

**5.2 Select the Appropriate Data-mining Method Based on Discussion**

Based on the above discussion, I think the best method is regression. The reason is that in our dataset, most attributes are numbers, and our ultimate goal is to predict an exact number. Instead of selecting in the ABC option. So, expression is the most appropriate method

**Step6: Data-mining Algorithm Selection.**

**6.1 Conduct Exploratory Analysis and Discuss**

In the data mining, there are many different kinds of algorithm and the following, we are going to talk some of them.

Decision tree is one of the famous tree modelers. It includes root node, internal node and leaf node. For the root node, it complete set with samples. Internal node is corresponding feature attribute test and for leaf node, it represents the result of decision. During the prediction, a certain attribute value is used at the internal node of the tree to determine which branch node to enter according to the judgment result until it reaches the leaf node to get the classification result. This is a supervised learning algorithm based on if then else rules. These rules of the decision tree are obtained through training, rather than manual formulation (Utgoff, Berkman & Clouse, 1997). Decision tree is the simplest machine learning algorithm. It is easy to implement, has strong interpretability, fully conforms to human intuitive thinking, and has a wide range of applications.

Another one is XGBoost. This is a simply called of “Extreme Gradient Boosting”, which is one of the kinds of tree modeler. This algorithm is suitable for forecasting the results from various factors (Chen & Guestrin, 2016).

Another kind of algorithm is linear regression, which originally a statistical concept, is now often used in machine learning. If there is a "linear relationship" between two or more variables, we can find out the "routine" between variables through historical data and build an effective model to predict the future variable results (Montgomery, Peck & Vining,2012).

Ridge regression is also a kind of linear regression, which is by giving up the unbiasedness of the least square method, a more practical and reliable regression method is obtained at the cost of losing part of the information and reducing the precision (Marquardt & Snee, 1975).

* 1. **Select Data-mining Algorithms Based on Discussion**

From the above table, we prefer to use linear regression be the model. The reason is that: these tree modelers seem more suitable for classification, instead of regression.

So, from the above discussion, linear regression seems to be the best modeler we could use.

**6.3 Build/Select Appropriate Model and Choose Relevant Parameter**

So, after decided to use linear regression, we are going to start use PySpark to achieve our goal. we could choose the parameter. They are PM10, CO, NO2, TEMP, Month, SO2 and DEWP. About wd and No, we will not be going to consider it because compared with other, their importance is so limit and no help on calculating the value of pm2.5. so, the output is PM2.5.

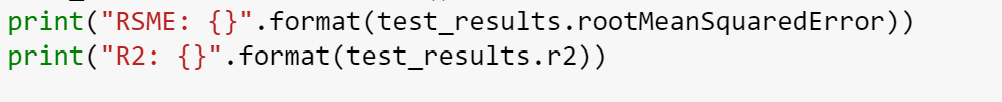
The original dataset includes multiple variables like hour, temperature, wind direction, wind speed, pressure etc. We divide the original dataset into 80% training set and 20% test set. We take temperature, pressure and wind direction as input, and take PM2.5 as output. After the regression process, we get the predictions PM2.5. We compare the predictive PM2.5 with the reality PM2.5 and get the mean square error. Good prediction result will be conducive to a deeper understanding of the causes, trends and related impacts of PM2.5.



**Step7: Data Mining**

**7.1 Create and justify test designs**

From the above discussion, we have decided to use linear regression model as our first choose. So, for the test, we will calculate the RMSE (Root Mean Square Error). It Is the square root of the ratio of the square sum of the deviation between the observed value and the true value and the observation times M. this value is used to measure the deviation between the observed value and the true value (Chai & Draxler, 2014).



**7.2 Conduct Data Mining**

In this step, I use linear regression in PySpark to build the modeler in the jupyter. At the end, we also print a picture shows our predict value and the true value.

A picture containing knife

Description automatically generated

**7.3 Search for Patterns**

The predict-real value shows like the following.

手机屏幕截图

描述已自动生成

The how much for each factor effect pm2.5 is that:

图片包含 游戏机, 截图

描述已自动生成

**Step8: Interpretation**

**8.1 Study and Discuss the Mined Patterns**

From the above pictures, we could see that pm10 always plays the most important role in prediction. The reason is because of the distinguish between these two identifications. The pm2.5 identified in Wikipedia is that the diameter of inhalable particles in the air is greater than 2.5 and pm10’s value is greater than 10 （Marcazzan et al., 2001）. So that means pm2.5 is a part of pm10. If pm10 stays at the high level, the pm2.5 perhaps be in a low level. There is some relationship between these two parts, but they could not totally decide another one’s value.

**8.2 Visualize the Data, Results, Models and Patterns**

By using PyShark, we could firstly draw a picture show how much for each factor effect pm2.5. the result is that:

图片包含 游戏机, 截图

描述已自动生成

And then, by testing our model, we also could draw the picture to show our predict pm2.5 from using linear regression and the real one.

手机屏幕截图

描述已自动生成

Also, in order to test our model, we calculate RMSE and show accuracy for training set and test set. The results are 0.5572 and 0.4525 separately.

**8.3 Interpret the Results, Models, and Patterns**

From the above picture, we could find that linear model has the smallest mean error (0.5572 training set accuracy, 0.4525 test set accuracy). It because these two models are based on regression analysis, which is more suitable for dealing with numbers (Darlington & Hayes, 2016), like our dataset.

Another kind of model is tree, like random tree, C&R tree, XGBoost tree. These tree models are more suitable for classification analysis, which already has some determined label. However, some tree model could be used to analyze regression number (Tso & Yau, 2007), for example, for the above picture, we could see that actually, C&R tree has the best preformation. The mean error is only 0.08.

From the patterns we could see pm2.5, CO, DEWP，NO2, SO2 and TEMP play important role in predicting pm2.5. However, NO2, NO and NOX also effect pm2.5(Christine, Peter&Linda, 2015).

**8.4 Assess and Evaluate Results, Models and Patterns**

After running our code, we will get the accuracy for training set and test set. They are 0.5572 and 0.4525. actually, these results are not very accurate, while we still would learn the around value of pm2.5 from the analysis. Moreover, for the picture, we also could find out that there are some blue dots above our red dots. That means our models not much suitable for analyzing pm2.5. however, compared with 30,000 attributes, the number of these blue dots are not occupied much. So, our models seem not too bad.

With increasing the size of training set, maybe our model will become much better. Or, we could find some other algorithm, which are more suitable for these kinds of dataset.

**8.5 Iterate Prior Steps(1-7) As Required.**

For this project, the first step is business understanding. In this process, we described our purpose of our research and success criterial, and then we introduce our dataset. It comes from the Machin Learning Repository website and we could use it freely. During our research, it is no doubt that we would face some risks, such as COVID-19. After having the roughly look at the dataset, we could find that this is regression and decide our data mining object. Firstly, we need to know how much for each factor impact the value of pm2.5, and then we need to find a suitable modeler for them. Finally, we make a plan for our project. It is almost a week.

For the data understanding step, firstly, we download our dataset from the website, and understand this dataset very well, like what the type for each column. With a roughly looking at the dataset, we also would do some hypotheses, like pm10 plays an important role in analyzing pm2.5 and so one. Moreover, it is not hard to realize that there are many missing values in it.

During selecting data, we deleted some unhelpful columns, like No., and station. Then, we use PySpark to calculate how many missing values for each column. However, the results show that they perform good with no more than 5%, which the highest number of missing data is 1776. So, we will keep the rest of them.

For step 4, we delete some rows if their pm2.5 are recorded by NA. after deleting these rows and column that we mentioned at above. We get our new clear dataset. Then, using python to create a new file and store them in it.

For the step 5, we talk about many deferent methods of data mining, and analyze that for our data, regression is our best choice. The reason is that our is to predict the level of pm2.5 and the level of pm2.5 is a continuous number.

For the step6, we need to consider our algorithm. Firstly, we study for four different algorithms, which are linear regression, linear regression and so one, and then based on these researches, we decided to use linear regression, because it seems the most suitable one.

For the step 7, we use PySpark to help us build our modeler. We divided our dataset for two part: one is for training and another one for testing. Moreover, we want to show the quality for our modeler, so we calculate RMSE as well. Therefore, for the results, we know how much for each factor impact pm2.5. We also get a picture about our predict value and the real value to see how much gap between our forecast and the reality.

**"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See:**[**https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.**](https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html)**).**

**I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."**

**Reference List**

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Utgoff, P. E., Berkman, N. C., & Clouse, J. A. (1997). Decision tree induction based on efficient tree restructuring. *Machine Learning*, *29*(1), 5-44.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, *29*(1), 3-20.

Kendrick, C. M., Koonce, P., & George, L. A. (2015). Diurnal and seasonal variations of NO, NO2 and PM2. 5 mass as a function of traffic volumes alongside an urban arterial. *Atmospheric Environment*, *122*, 133-141.

Fisher, R. A. (1920). 012: A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error.

Darlington, R. B., & Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Publications.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761-1768.

Marcazzan, G. M., Vaccaro, S., Valli, G., & Vecchi, R. (2001). Characterisation of PM10 and PM2. 5 particulate matter in the ambient air of Milan (Italy). *Atmospheric Environment*, *35*(27), 4639-4650.

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, *2*(2), 43-47.

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, *58*(304), 933-942.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*, *25*(2), 103-114.

Zheng, M., Salmon, L. G., Schauer, J. J., Zeng, L., Kiang, C. S., Zhang, Y., & Cass, G. R. (2005). Seasonal trends in PM2. 5 source contributions in Beijing, China. *Atmospheric Environment*, *39*(22), 3967-3976.