# Beijing PM2.5 Concentration Data Analysis, Modeling, and Visualization
# FA 20 COGS 109 Final Project Written Report

Jeffrey Feng, Beibei Du, Yijia He, Scott Yang

*University of California San Diego, Cognitive Science*

## 1a). Background: Data description

We selected one of the dataset marked for regression predictive tasks from the UCI machine learning repository: (https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data). The original data comes with 12 separate csv files, which are recordings of weather data of 12 different environment monitoring stations in Beijing from 2013 to 2017. Since the "station" column for each csv file is the same, as the first step of data processing, we concat all 12 csv files into one dataframe. The concatenated dataset has 18 columns, and 420768 rows. The columns are row number, year, month, day, hour, PM2.5 (ug/m^3), PM10 (ug/m^3), SO2 (ug/m^3), NO2 (ug/m^3), CO (ug/m^3), O3 (ug/m^3), temperature (°C), air pressure (hPa), dew point temperature (°C), rain precipitation (mm), wind direction, wind speed (m/s), station name. Station name, row number would not be used for any data analysis, since each csv file contains only one station name, and since all the stations are located in one city, it does not provide extra information if we aggregate the data later.

## 1b). Background: Predictive Task

A meaningful predictive task for this dataset is to estimate the amount of PM2.5 concentration in Beijing, if we only have access to the information of other variables like NO, SO2 concentration. Other than the use of forecast modelling, this predictive task can also be potentially useful for studying the relationship between the amount of PM2.5 with other weather conditions. Although we have abundant feature and samples, we could also predive the amount of PM2.5 based on the amount of PM10, because as both pollution particles in the air, they differ in diameter, if we assume that PM2.5 is harder to detect than PM10, it is useful to make models of PM2.5 based on the amount of PM10, if the future study have only access to devices that detect concentration larger particles. Further in this report, we will see that the concentration of PM10 would be a very important factor.

## 2a). Method: Data cleaning & wrangling

First of all, a large data with over 400000 samples may be unnecessary for the analysis because the original data was recorded in a high time resolution: the PM2.5 data was sampled every hour at each station. Furthermore, for 12 columns, there are many missing values and the number of missingness ranges from 300 to 20000. A method to overcome these is to group the dataset over each day, and aggregate all the other columns over the mean per day. We will groupby(['year', 'month', 'day']).mean().reset_index(). As a result, we lost the wind direction column as it is the only categorical feature, but we produced a new dataframe with 1461 rows, each representing a day starting from 3/1/2013 to 2/28/2017, shown in Figure 1.

| | year | month | day | No | hour | PM2.5 | PM10 | SO2 | NO2 |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 2013 | 3 | 1 | 12.5 | 11.5 | 7.326389 | 12.255245 | 9.280142 | 21.405738 |
| **1** | 2013 | 3 | 2 | 36.5 | 11.5 | 31.475694 | 40.616725 | 32.007989 | 56.704889 |
| **2** | 2013 | 3 | 3 | 60.5 | 11.5 | 79.291667 | 111.104167 | 49.386760 | 77.021429 |
| **3** | 2013 | 3 | 4 | 84.5 | 11.5 | 21.731449 | 40.601399 | 18.805865 | 43.134273 |
| **4** | 2013 | 3 | 5 | 108.5 | 11.5 | 132.439114 | 159.236111 | 71.333333 | 104.256506 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1456** | 2017 | 2 | 24 | 34956.5 | 11.5 | 25.286713 | 38.472028 | 9.534965 | 44.614286 |
| **1457** | 2017 | 2 | 25 | 34980.5 | 11.5 | 11.392226 | 21.583039 | 5.590106 | 30.402827 |
| **1458** | 2017 | 2 | 26 | 35004.5 | 11.5 | 27.785965 | 45.066667 | 10.021053 | 50.463158 |
| **1459** | 2017 | 2 | 27 | 35028.5 | 11.5 | 66.804511 | 97.183521 | 16.569811 | 76.162264 |
| **1460** | 2017 | 2 | 28 | 35052.5 | 11.5 | 14.945848 | 28.853047 | 6.448905 | 32.700730 |

1461 rows × 16 columns

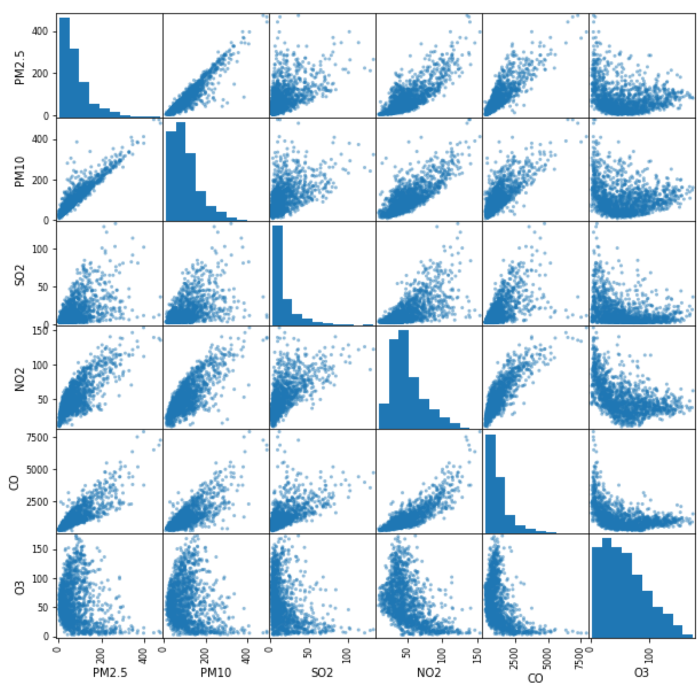
(Fig 1. Transformed dataset. Only a few of the 16 columns are shown here. ‘PM2.5’ is the dependent variable)

## 2a). Method: Data visualization & Model selection

The primary technique for this project is the linear regression, that makes prediction on one of the continuous variables, namely the PM2.5 concentration. Exploratory analysis by plotting can be helpful when we try to make useful feature extraction to optimize the model by selecting associated columns that we can find a strong relationship with the predictive label.

Without expertise knowledge, we are not certain if the association between time variables and the amount of PM2.5 has a strong pattern. However, intuitively, we can make guesses and believe that the following columns --- 'PM2.5', 'PM10', 'SO2', 'NO2','CO', 'O3' --- that are all weather variables, might have some obvious linear relationship with each other.

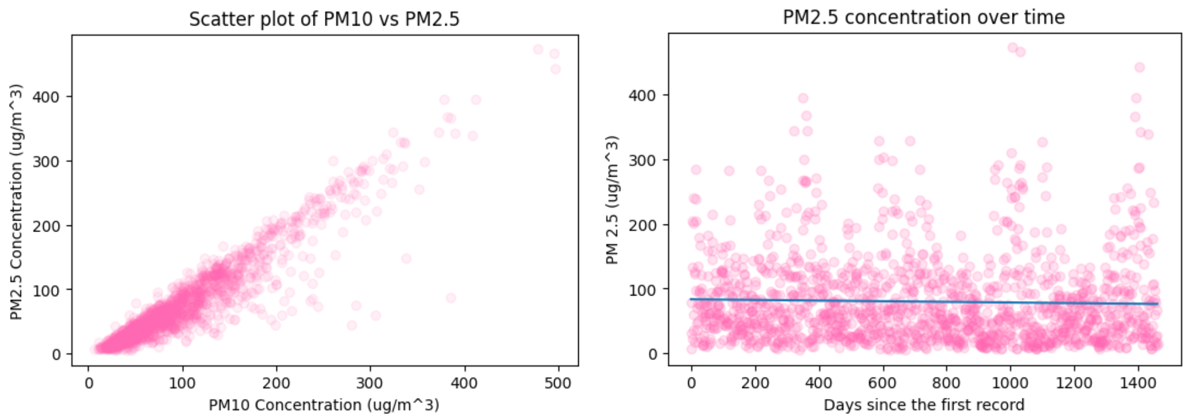
By plotting out the scatter matrix of the above selected features, we can see which two variables might have a correlation with the other one, and the histogram distribution of these variables on the diagonal line. Based on the output of the scatter matrix, we decided to have prsa['PM2.5'] as our dependent variable(y) such that we shuffle the dataset by using the df.sample() and select the first 1200 entries as our Y trained, and the entries after 1200 as our Y-tested data.




(Fig 2. Scatter matrix of PM2.5, PM10, SO2, NO2, CO, and O3 concentration)

There are interesting findings from the scatter matrix: the O3 concentration seems to have a negative correlation with all the other variables, while the other variables all show some degree of correlation. The shapes of these correlations look like a fan, indicating that it is hard to predict the y when the given x increases. All these plots seem to have similar correlation coefficients, except O3, and the scatter plot of PM10 and PM2.5 (shown in a larger size in Fig3), suggested stronger association. We then infer that it is reasonable to make a **baseline model** using the amount of PM10 alone to predict PM2.5.

What about other variables, like the time variables (Day, Month, Year)? It is reasonable to use the model PM2.5=w0+w1*days from the first data records, because there must be up and down in the amount of PM2.5 in the past. Shown in Fig 4, From the blue line, it indicates that there is no linear relationship between the days and the concentration of PM2.5. This pretty flat line does not provide any positive nor negative correlation as we thought it might have a trend (As the days increase, there might be a decrease in the PM2.5 concentration). The pink dots scatter around various values of PM2.5 and it does not tend to increase nor decrease despite there being few extreme high values at days around 400, 1000 and 1400. Despite the Chinese government starting to take action on protecting the air quality and concerns the environmental issues, it does not seem to reflect on ameliorating the air quality and the amount of the PM 2.5. As the action plan of Chinese government requires, they initiate “strengthening industrial emission standards, phasing out small and polluting factories, phasing out outdated industrial capacities, upgrades on industrial boilers, promoting clean fuels in the residential sector, and strengthening vehicle emission standards”(Zhang et al., 2019). However, from what our resulting visualization reveals, there is no decreasing in the PM 2.5 as the time passes. Based on the result, we can further deduce that there is no significant effect on executing these action plans on protecting the environment in the time frame that we are interested in. Although one can still argue that there were up and down in Fig 4, especially with the spikes at days = 400, 700, 1000, and 1400, is it not the focus of this study to further explore the reason.




(Fig. 3 & 4 PM2.5 scatter plot with PM10 and days since the first record on x axis)

## 2b). Method: Feature Selection

From the plot we decided to create a linear regression model with degree = 1 on various columns. The main question is, how do we find the best features combinations, and how do we define best?

There are 15 features/columns, we decide to find out for each N number of columns (which range from 1 to 15, if we do not count the biased term that we also feed the linear regression model), which N columns perform the best. Selecting N columns from 15 columns will generate 15! / N! * (15 - N)! combinations.

For example, if N = 2 (using 2 features), we will find all the permutations (order does not matter, therefore no repetition) of features, such as concentration of PM10 + NO, or Day + Year, create a train set with only columns and the biased term, find the weights of the linear regression through np.linalg.lstsq(), and calculate the R^2 values, or the coefficient of determination, (the closer the value to 1, the better the model is), and the mean squared error (the less this value is, the better the model is). R^2 value and mean squared error will be the performance metrics that we use, to find the best combination.

Out of 15!/(2! * 13!) = 105 combinations shown in Fig 5., we will select the best one:

| | Features Used | R^2 | Mean Square Error |
|---|---|---|---|
| 62 | [PM10, CO] | 0.922751 | 378.788957 |
| 68 | [PM10, WSPM] | 0.903191 | 474.697564 |
| 61 | [PM10, NO2] | 0.898105 | 499.636819 |
| 63 | [PM10, O3] | 0.889272 | 542.951964 |
| 65 | [PM10, PRES] | 0.887406 | 552.101289 |
| ... | ... | ... | ... |
| 57 | [hour, DEWP] | 0.000861 | 4899.238153 |
| 36 | [day, DEWP] | 0.000610 | 4900.467438 |
| 28 | [day, hour] | 0.000070 | 4903.118056 |
| 14 | [month, day] | -0.000737 | 4907.072496 |
| 16 | [month, hour] | -0.000800 | 4907.384297 |

105 rows × 3 columns

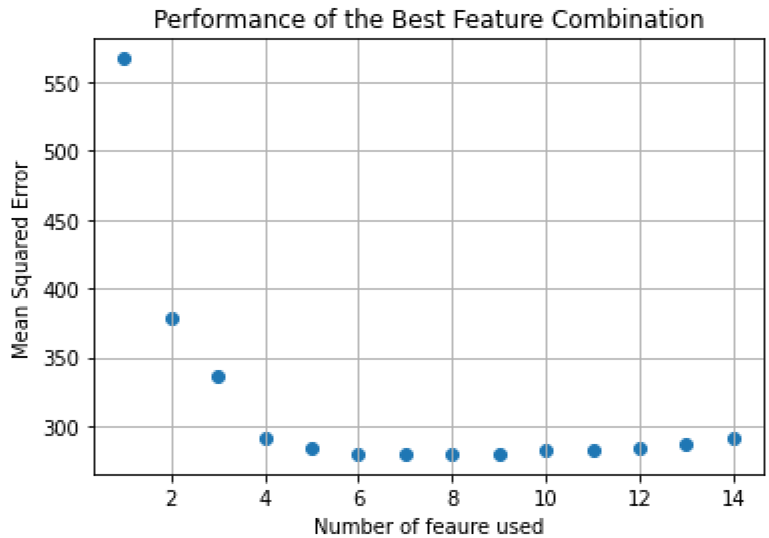
(Fig 5. N = 2 feature selection)

## 3). Results

For N = 1 (baseline model) to N = 15, we ran the same analysis that we previously showed, and we compared the best feature combination at each N shown in Figure 6 and 7 below.

| | Number of feaure used | Best Feature Combination | R^2 Score | Mean Squared Error |
|---|---|---|---|---|
| 1 | 1 | [PM10] | 0.884412 | 566.779301 |
| 2 | 2 | [PM10, CO] | 0.922751 | 378.788957 |
| 3 | 3 | [PM10, CO, DEWP] | 0.931512 | 335.828321 |
| 4 | 4 | [PM10, CO, TEMP, DEWP] | 0.940624 | 291.150063 |
| 5 | 5 | [month, PM10, CO, TEMP, DEWP] | 0.941982 | 284.487289 |
| 6 | 6 | [month, PM10, CO, TEMP, PRES, DEWP] | 0.942853 | 280.219598 |
| 7 | 7 | [month, PM10, CO, TEMP, PRES, DEWP, RAIN] | 0.942941 | 279.784157 |
| 8 | 8 | [month, hour, PM10, CO, TEMP, PRES, DEWP, RAIN] | 0.942941 | 279.784157 |
| 9 | 9 | [year, month, hour, PM10, CO, TEMP, PRES, DEWP... | 0.942900 | 279.985417 |
| 10 | 10 | [month, hour, PM10, CO, O3, TEMP, PRES, DEWP, ... | 0.942359 | 282.640932 |
| 11 | 11 | [year, month, hour, PM10, CO, O3, TEMP, PRES, ... | 0.942259 | 283.131583 |
| 12 | 12 | [year, month, hour, PM10, NO2, CO, O3, TEMP, P... | 0.942139 | 283.716920 |
| 13 | 13 | [year, month, hour, PM10, SO2, NO2, CO, O3, TE... | 0.941411 | 287.287837 |
| 14 | 14 | [year, month, day, No, hour, PM10, NO2, CO, O3... | 0.940560 | 291.463616 |

(Fig 6. Performance of best feature selection combinations of different number of features)




(Fig 7. Number of features vs. Best performance)

## 4). Discussion

At N = 7 and 8, we get the same MSE and R^2 values. For the simplicity of real life practice, we omit the extra one feature, and conclude from the result plots shown above, that the linear regression model attain its best performance when we use [month, PM10, CO, TEMP, PRES, DEWP, RAIN], with the R^2 score as high as 0.942941 and the minimal mean squared error as 279.78. If we increase the number of features furthermore, the performance decreases.

This result is not surprising because our experience of living in China and reading news on weather in the capital has informed us that PM2.5 can be high in some specific month due to variation in industry output and amount of rain and wind. As a result, month is actually the best feature we got, even though we have found that the number of days does not contribute to our confidence in the PM2.5 prediction (fig 4). One of our initial concerns was that due to similarity to PM2.5, PM10 will dominate the feature input and cause other features to add variance to the model, but our results shows that our model can significantly beat the baseline model (R score = 0.88) and make much more accurate predictions.

## 5). Reference


Data Source: Song Xi Chen, csx '@' gsm.pku.edu.cn, Guanghua School of Management, Center for Statistical Science, Peking University. https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data#

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., . . . Hao, J. (2019, December 03). Drivers of improved PM2.5 air quality in China from 2013 to 2017. Retrieved December 17, 2020, from https://www.pnas.org/content/116/49/24463