Project Proposal

Paul Holaway (paulch2), Albert Li (xiangl9)

October 21st, 2022

Data Description

Our data is a collection of readings of PM2.5 in the city of Beijing, China. It was retrieved from Kaggle as a part of a collection of PM2.5 readings for five major Chinese cities (Beijing, Chengdu, Guangzhou, Shanghai, and Shenyang). We chose to use just Beijing to train our models as it is the capital city of China. The data collected was between January 1st, 2010 and December 31st, 2015. However, not all recording sites began recording at the same time. There are 946,512 total data inputs and 91124 `NAs` out of them, accounting for nearly 10% of the overall data set, mainly because there are no PM2.5 data recorded for all four PM stations until 8 A.M. Mar 5th, 2013. We plan to build a model for at least two of these recording stations to compare how recording at a different location can effect the best fit model. A description of the data is below.

Data Set Variable Descriptions

- `No`: Row Number
- `year`: Year when the data was recorded in the row.
- `month`: Month when the data was recorded in the row.
- `day`: Day when the data was recorded in the row.
- `hour`: Hour when the data was recorded in the row (in 24 hour format).
- `season`: Season when the data was recorded in the row.
  - 1 = December - February
  - 2 = March - May
  - 3 = June - August
  - 4 = September - November
- `PM_Dongsi`: PM2.5 concentration ($\mu g/m^3$) recorded at Dongsi, Beijing (52% `NAs`).
- `PM_Dongsihuan`: PM2.5 concentration ($\mu g/m^3$) recorded at Dongsihuan, Beijing (61% `NAs`).
- `PM_Nongzhanguan`: PM2.5 concentration ($\mu g/m^3$) recorded at Nongzhanguan, Beijing (53% `NAs`).
- `PM_US.Post`: PM2.5 concentration ($\mu g/m^3$) recorded by the US recording station at Dongsi, Beijing (3% `NAs`).
- `DEWP`: Dew Point ($^\circ C$)
- `TEMP`: Temperature ($^\circ C$)
- `HUMI`: Humidity (%)
- `PRES`: Pressure (hPa)
- `cbwd`: Combined Wind Direction
- `Iws`: Accumulated wind speed (m/s)
- `pressure`: Hourly Precipitation (mm)
- `Iprec`: Accumulated Precipitation (mm)

Subsetting

We then made four sub-data sets (`Dongsi`, `Dongsihuan`, `Nongzhanguan`, and `USPost`) and for each we:

- Removed all `NAs` from the PM column corresponding to the location it was recorded in, as PM2.5 is the response variable and the predictors would not be valuable without this.
- Created one column for each sub-data set that contains all time-related information by combining the year, month, day, and hour columns. All other ten columns from the main data set are included as well. Those remaining columns are as follows.
- `Recorded`: The exact hour, day, month, and year when the data was recorded in the row.
- `season`: Season when the data was recorded in the row.
  - 1 = December - February
  - 2 = March - May
  - 3 = June - August
  - 4 = September - November
- `PM_...`: PM2.5 concentration ($\mu g/m^3$) recorded at the specific location in Beijing.
- `DEWP`: Dew Point ($°C$)
- `TEMP`: Temperature ($°C$)
- `HUMI`: Humidity (%)
- `PRES`: Pressure (hPa)
- `cbwd`: Combined Wind Direction
- `Iws`: Accumulated wind speed (m/s)
- `pressure`: Hourly Precipitation (mm)
- `Iprec`: Accumulated Precipitation (mm)

Questions To Address

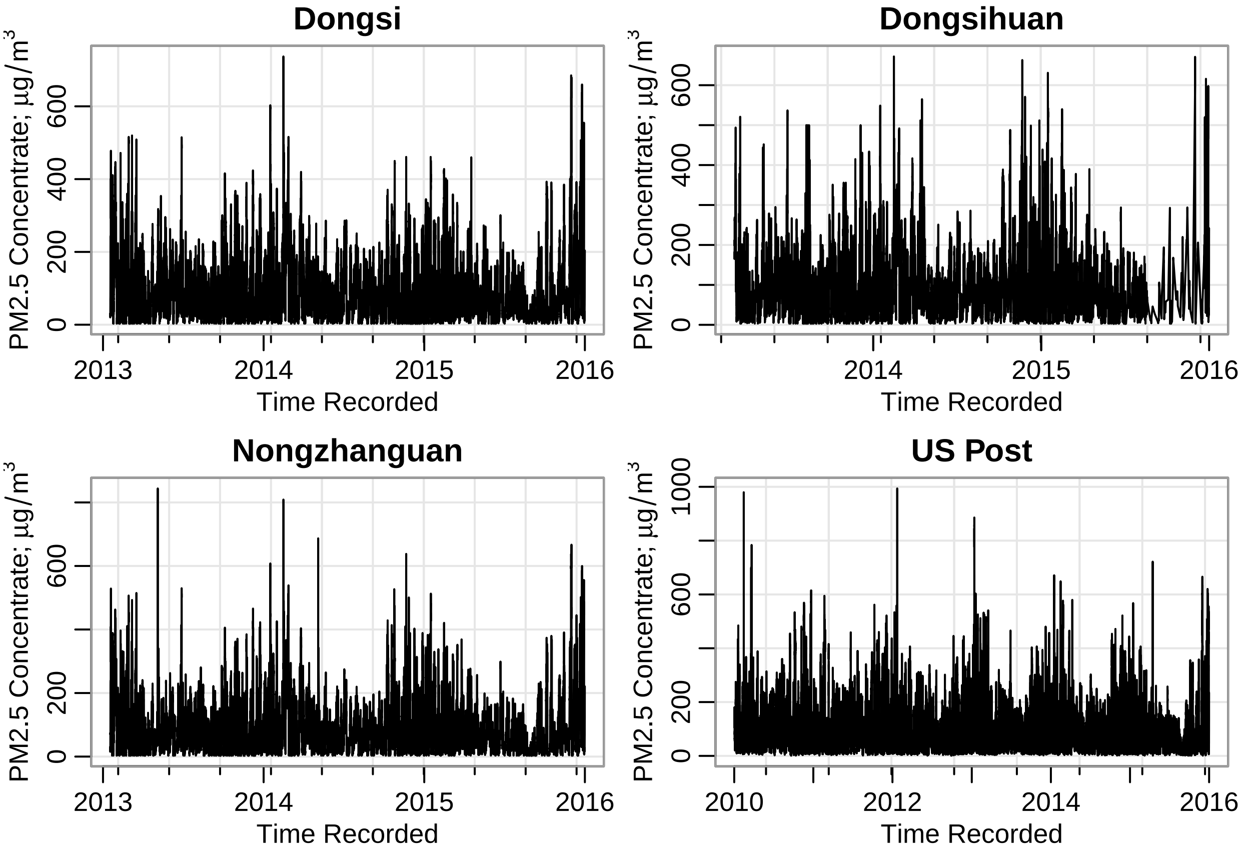
1. What is the seasonal pattern for PM2.5 in these locations?

- Are they all the same or are some different than others?

2. How does a different recording station location effect the best fit model and therefore predictions?
3. What will the predictions for the PM2.5 levels be?

- Is the amount getting larger or smaller?
- We hope smaller, but would like to know if it is not.

Visualization



Plans For Analysis

SARIMA Model

For this part, we plan to look for a specific seasonal trend as there appears to be one present in all four of the recordings above. Due range and volatile changes in the value of PM2.5, we believe a linear regression model would not yield meaningful results. We will investigate using transformations and differencing as needed once we begin our analysis. A log-transformation could be useful here as we have large variation across the data. We will also figure out the seasonal trend(s) for the data.

Advanced Topic: ARCH/GARCH Model

For this part, we plan to look at building an ARCH and/or GARCH model for the recording locations that we analyze in the first part. This is because when looking at the time plots for the data, the PM2.5 shows signs of heteroskedasticity which would be best solved using a ARCH or GARCH model. We then plan to compare the effectiveness of both models and see which one predicts better.