ISYE 6416 Project Proposal

Beijing PM2.5 Time Series Analysis and Prediction using Regression and Markov Model at Different Time Scales

Mengmeng Liu, Xin Cao

1. Problem Statement

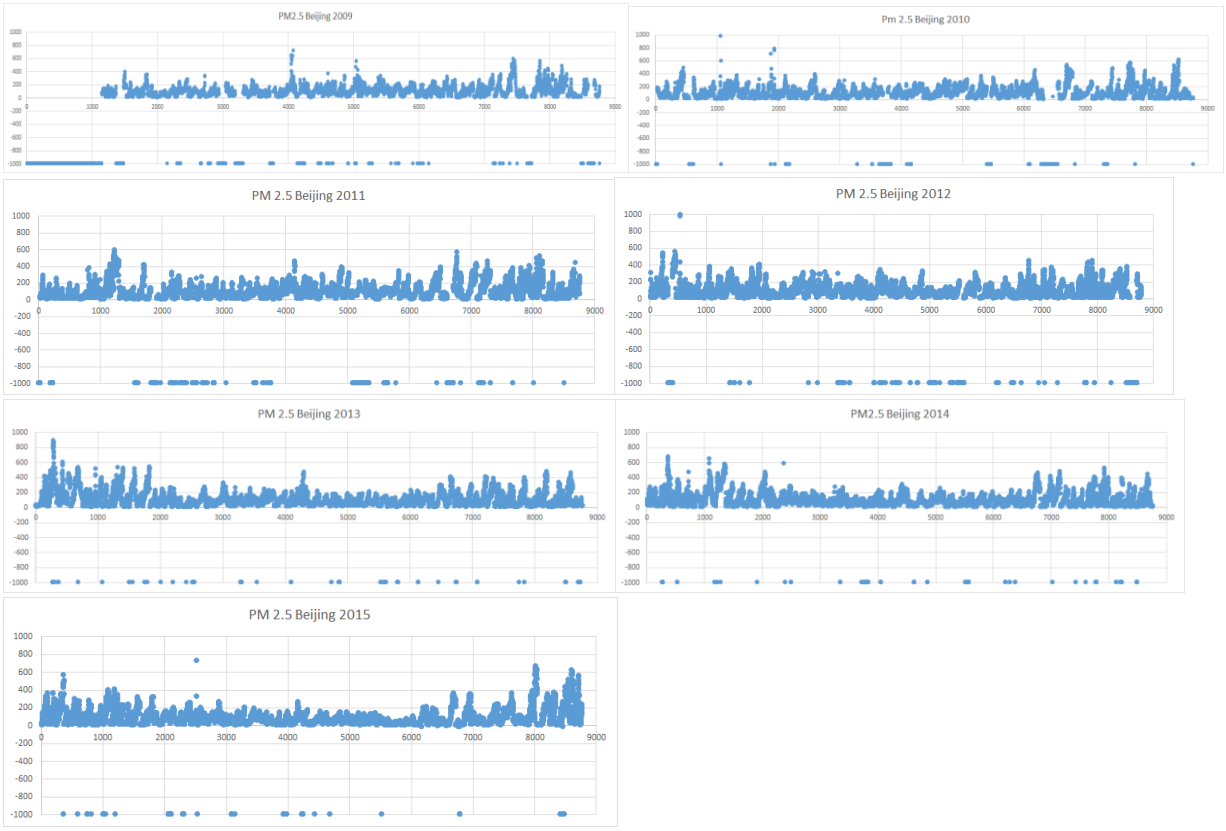
The air pollution in modern cities is a severe problem which significantly affects human's life and health. PM 2.5 is a measurement a type of particulates or aerosol with a scale size less than 2.5 micrometers which usually suspends in the atmosphere. This majority of this aerosol consists of some chemicals such as organics, sulphate, amine, nitrate, black carbon and so on. The cause of PM 2.5 is very complex because it is sourced not only from the protosomatic emission or production, but also from secondary emission or production. The protosomatic production includes but not limited to vehicles emissions, power plants emissions or even natural fires. The secondary production mainly comes from varieties of chemical reactions between different chemicals in the atmosphere, whose process is usually very complicated to investigate. Besides, the physical condition of atmosphere is also an important factor to affect PM 2.5, such as the temperature, pressure, humidity, wind orientation, wind speed. And therefore many atmospheric scientist developed some chemical model to explain and predict the influence of these chemical reactions on the production of PM 2.5.

However, even the models have been developed better and better, the prediction of PM 2.5 is still a hard problem. Because it is not only related to the chemical reactions, physical parameters, but might be also affected by some unknown minor factors such as human local activities which are hard to accurately measured and predicted. On the other hand, no matter which model of chemical or physical is used to study and predict PM 2.5, they more or less made some assumptions in the models because it would become extremely complicated and nearly impossible to solve without these assumptions. But these assumptions might also cause some unpredicted errors since the nonlinear behaviors could result in chaos. An alternative method is to use statistical models to investigate the variation of the PM 2.5 so that we do need to consider the complex intermediate process. However, there are still some unknown factors not completely included in the statistical model, an explanation is that these factors might be tolerated in the confidence interval and confidence level. The details of the models will be discussed at next section.

2. Data Source

The data is download from “U.S. Department of State Data Use Statement”, its website is : http://www.stateair.net/web/post/1/1.html . The data include the **hourly** PM 2.5 observation data from 2009 to 2016 in Beijing.

The scatter plot of PM 2.5 in each year are shown in following figures. There are some missing data in the data set, which are marked as “-999”. So we need clean the data before use them to train our model.



3. Methodology

3.1 Models

Two different models are considered to be used to do the time series analysis of PM2.5 in Beijing during 2009-2015. After obtained the model, we use them to predict the PM 2.5 in 2016, and compare our prediction with the real data.

- **Polynomial Regression**
  Build a polynomial regression model to describe the relationship of PM2.5 with time in Beijing. The independent variable is time, and the dependent variable is the value of PM 2.5 in unit of $\mu g/mg^3$,

- **Markov Chain Model (MCM)**
  In MCM model, we solve the following problem:

Given an observation sequence O, and a space of possible models, how do we adjust the parameters *u*, which include transition probability.

In the project, MCM is used to train the model with past historical data, which is the PM 2.5 during year 2009-2015. After the training, we can then use the model to perform further prediction (eg: predict the PM2.5 in 2016).
the observation sequence O will be : sequence value of PM2.5,

Note that by using MCM, we assume that the PM 2.5 in time *t+1* only depends on the PM2.5 in time *t*.

3.2 Analyze the sensitivity of model

We will train our model at different time scales, including hourly, monthly, seasonally, and yearly. And then analyze how our prediction result will change with time scales.

4. Expected Results

Expected Results are models that can be used to predict the PM 2.5 in future.
Same model may give different prediction results at different time scales. The hourly scale may make the model be easily overfitting, while yearly scale let the prediction under estimated.

5. Reference

[1] Yingjian Zhang. 2001. prediction of financial time series with hidden markov models. Shandong University, China.
[2] https://en.wikipedia.org/wiki/Markov_chain#Transience
[3] https://en.wikipedia.org/wiki/Particulates#Sources_of_atmospheric_particulate_matter