STAT 420 Project Report 1

Winston Chen, Daniel Hooks, Naidan Ganbold, Erik Janc, Michael Lee

October 19, 2021

Dataset

Link to dataset: https://www.kaggle.com/sogun3/uspollution
The U.S. Air Pollution Data includes daily measurements of four different pollutants across various states, cities, and counties across the country from 2000 to 2016. The data includes the following attributes:

- State Code
- County Code
- Site Num
- Address
- **State**
- **County**
- **City**
- **Date Local**
- **[pollutant] Units**
- **[pollutant] Mean**
- [pollutant] Air Quality Index (AQI)
- [pollutant] 1st Max Value
- [pollutant] 1st Max Hour

The relevant attributes we plan to incorporate into our analysis are bolded. The four pollutants are Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Carbon Monoxide (CO), and Ozone (O3).

There are over 1.4 million records in this dataset. We plan to aggregate the daily data into monthly data, and consider only the following cities (tentatively, subject to change):

- Los Angeles
- Chicago
- New York City
- Denver

Research Objectives

We aim to consider select major cities to compare and contrast their pollution data and identify trends that relate them. We also plan to consider potential external events that may influence what we see in each time series, such as natural disasters (forest fires), pollution laws, etc. Further, we will attempt to forecast future pollution levels for each city and predict how we expect future pollution levels in each city to compare with each other.

Statistical Models and Methods

- Moving Average
- Smoothing
- Autoregressive Moving Average
- Maximum Likelihood

Group Work and Timeline

We will use GitHub in order to facilitate efficient collaboration on a centralized R Notebook for this project.

- Link to GitHub repository: https://github.com/michaeldlee23/stat420-project

We plan to have our group work together on the majority of the project, meeting at least once per week for status updates.

- Week of 10/18: Data exploration and preprocessing
    - Get data into a usable format with only relevant rows and data
    - Obtain data visualizations (time series plots, moving averages)
- Week of 10/25: Begin forecasting analysis
- Week of 11/1: Continue forecasting analysis
- Week of 11/8: Record initial findings and tentative conclusions, finish and submit the second report
- Week of 11/15: Make any necessary changes to models or additional analysis, finish and submit the final paper