ISYE-6414 Final Report - Air Quality Index

Ishika Arora, Ian Jiang, Tess Leggio

November 30, 2018

Problem Statement

Air quality is a factor that is cited as both contributing to and being affected by climate change (1), and has been shown to have a direct impact on human health. As a factor with such important consequences, our team decided to study factors that impact air quality. For our project, we focus on answering the following question: *How is air quality across the United States related to demographic, emissions, and weather factors?* Specifically, we chose to investigate the impact of factors related to economic output and climate, including temperature, population, per capita income, greenhouse gas emissions, and industrial power consumption and generation. Our response variable is the Air Quality Index (AQI), a measure of local air quality calculated by the Environmental Protection Agency (EPA) based on the concentrations of 6 major classes of ground-level pollutants: ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, and two measures of particulate matter (PM2.5 and PM10).

The scope of our analysis ultimately included county-level data for all US counties with available AQI data. To limit the size of our dataset to a size manageable within R, we used daily data from 2016, the most recent year with data available for each of the factors included in our analysis.

(1): https://www.epa.gov/air-research/air-quality-and-climate-change-research

Data Collection

Most of the data we used in our analysis is available from government websites. Data is available for download as CSV files from the following sources: per capita income data from bea.gov, population data from census.gov, powerplant energy consumption and generation data from eia.gov, and AQI data from aqs.epa.gov.

Weather data were significantly more challenging to collect. Daily weather sensor data is available from aqs.epa.gov, however the county data is far less complete than AQI data, so to avoid omitting a large portion of data from analysis, which could introduce bias into the model we instead collected temperature data from api.mesowest.net, an API which allows straight-forward programmatic extraction of data. By setting parameters in the URL, daily temperature and other weather averages can be scraped for any given day, state and county. We exploit this feature by looping through all counties in the AQI data and all days within our date range, sending a "get" request for each iteration. Data is available in JSON format, and includes daily averages for all stations within a given county. Therefore, a global measure of temperature and other weather factors is easily calculated by converting the JSON format into a Python dictionary and averaging across all stations. The speed of the script is highly dependent on local factors and internet speed, but on average it takes about 10 minutes per county for each of the 1053 counties included in our analysis.

Data Cleaning

Upon collecting data from each of the 6 identified data sources, we were faced with the task of merging the datasets together. First and foremost, we had to ensure that the datasets merged together well. We had to merge data based on state name, county name, year, month, and date, so it was essential that these fields were in the exact same format across data sources to avoid loss of data. In particular, we spent much time ensuring that county names were in the proper format. Since county name is a string field, it was not

uncommon to see several different formats (for example, Saint John vs. St. John vs. StJohn). To avoid having several sparsely populated observations for each county, we wrote python scripts to format each dataset to use the same format and merge the datasets together.

After merging the datasets, we noted that there were several columns that were only sparsely populated, especially within weather data such as wind speed, relative humidity, and precipitation. As these data were sparse, we decided not to use them in our analysis because it may have introduced bias into the model. As such, we dropped the columns from our dataset.

Next, we checked our merged data to ensure data quality.

- C) Dealing with missing data
- D) Checking for data quality
- E) Rescale the predictors/response as necessary *** we should do this. see p. 103 of textbook ***
- F) Choosing subset of data to enable us to work within R