Project Proposal: Air Quality Prediction Based on Regions and Seasons

Julianne Lin and Thitaree Tanprasrt

1. Problem

For this project, we would like to use big data techniques to extract information and make prediction from data in a health-related field. Based on an initial search through the governmental database, we would like to explore the data related to air quality. More specifically, we would to explore data of concentration of fine particulate matters. We would like to use regression techniques to predict air quality based on information about regions (location) and/or by seasons (time), since the data for these categories are readily available.

2. Datasets

Our data comes from data.gov, specifically the health-data catalog from the site, which contains national data made public by the federal government. A few of the data sets under consideration right now include:

- **Daily Census Tract-Level PM2.5 Concentrations:** this dataset is used by the CDC's National Environmental Public Health Tracking Network. It is separated into 3 datasets: year 2001-2005, year 2006-2010, year 2011-2014.
  Links:
  - https://catalog.data.gov/dataset/daily-census-tract-level-pm2-5-concentrations-2001-2005
  - https://catalog.data.gov/dataset/daily-census-tract-level-pm2-5-concentrations-2006-2010
  - https://catalog.data.gov/dataset/daily-census-tract-level-pm2-5-concentrations-2011-2014
- **Air Quality Measures on the National Environmental Health Tracking Network:** this dataset is provided by the Environmental Protection Agency (EPA). It contains ozone concentration and PM2.5 concentration data collected from over 4000 monitoring stations.
  Link: https://catalog.data.gov/dataset/air-quality-measures-on-the-national-environmental-health-tracking-network
- **CDC WONDER: Daily Fine Particulate Matter:** this dataset contains measurements of PM2.5 concentration collected based on geographical regions during the year 2003-2008.
  Link: https://catalog.data.gov/dataset/cdc-wonder-daily-fine-particulate-matter-cc052

We have already downloaded the csv files for these and verified that the amount of data fulfills the requirement of around half-million data points.

3. Methods

Currently, we have limited knowledge about the regression techniques we can use for our prediction model. We plan to use the linear regression method, which we have learned recently in class, to predict the air quality based on information about seasonal period and coordinates of locations. Based on the accuracy of this method, we could also consider using classification model to gain more insights about seasons and regions, which are labeled data.