Exploratory Data Analysis and Forecasting of US and California Air Quality using SARIMAX and TBATS Time Series Approach

Angela Cao

Air pollution has been a persistent problem impacting citizens' health and the environment. The ability to model, predict, and monitor air quality is relevant especially in urban cities. Using Python, this study investigated sixteen years of US air quality data first through an exploratory data analysis and then focused on data from the state of California, building and comparing models for air quality forecasting. The exploratory data analysis identified general and interesting patterns. Multiple naïve and advanced time series forecasting methods provided a comprehensive forecast of the air quality of four air pollutants (SO_2, NO_2, O_3, CO_2).

It is found that over the sixteen years, the CO, SO_2, and NO_2 levels had downward trends while O_3 fluctuated at a high level. There was a 0.67 correlation between NO_2 and CO_2 levels in California, and these two pollutants were most correlated in the winters and least in the summers. Yearly seasonality was observed for each pollutant, with O_3 having a seasonal pattern opposite of the rest. Both advanced models of SARIMAX (the fourier terms approach) and TBATS fitted well with SARIMAX fitting better for NO_2 and O_3, and TBATS fitting better for CO and SO_2. In the upcoming summer, I plan to incorporate data of meteorological/human-caused parameters from the National Climate Data Center or California Transportations to compare the magnitudes of influences and improve the models. Predicting air quality and better understanding air pollution are complex tasks that require sustained scientific attention for the sake of humanity and environment.

Mentor: Mr. Johnny Ma; NYU Center of Data Science