Using Machine Learning to Analyze the Effects of Air Quality and Meteorological Exposures on Mortality

Office of UNDERGRADUATE RESEARCH
THE UNIVERSITY OF UTAH

Grace Liu[1]; Sejal Mistry, BA[2], Dr. Ramkiran Gouripeddi, MBBS, MS[2,3], Dr. Julio Facelli, PhD[2,3]

[1]Department of Atmospheric Sciences, University of Utah, [2]Department of Biomedical Informatics, University of Utah, [3]Clinical and Translational Science Institute, University of Utah

Background

- Air pollution has been linked to many adverse health outcomes and corresponding mortality rates. Climate is also thought to influence both air quality and mortality rates.
- However, the relationship between air quality, meteorological variables, and health outcomes is not well understood.
- Machine learning can be used to analyze the complex relationships between these variables, providing knowledge of how mortality is influenced by environmental conditions.

General process:

- Gather longitudinal data
- Choose the number of clusters k using the elbow method and internal evaluation metrics (Calinski-Harabasz, Davies-Bouldin Index, silhouette coefficient)
- Determine best clusters

Objective

We aim to use unsupervised temporal machine learning to analyze the relationship between environmental factors, air quality, and mortality data.

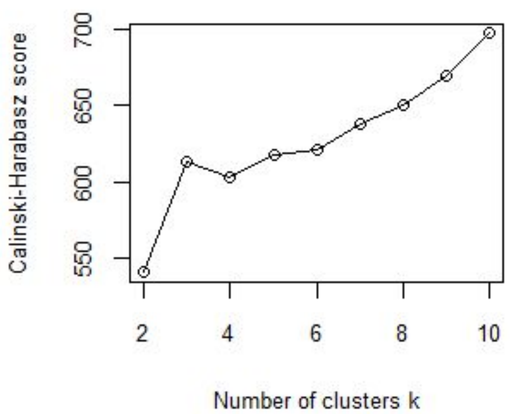
Data

- Daily measurements for 15 years
- 17,500 vs. 800 locations
- NO2 and O3 measured by OMI
  https://disc.gsfc.nasa.gov/datasets/OMI_MINDS_NO2d_1/summary?keywords=omi
- CO, PM2.5, SO2 measured by MERRA-2
  https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4/summary?keywords=merra%20pm2.5
- Wind, temperature, pressure, humidity data from the EPA
  https://aqs.epa.gov/aqsweb/airdata/download_files.html

Methods

- Previously explored large row anomalies in the NO2 and O3 data
- Compile longitudinal data in Python
- Cluster joint longitudinal trajectories of meteorological data using k-means for joint longitudinal data (kml3d) in R
- Evaluate cluster performance using the Calinski-Harabasz score and other metrics

Elbow method for 2020 temperature data



Conclusions

- There are many steps involved in processing the data for our variables used, such as considering missing values and anomalies at certain locations, the ratio of features to observations (aim for 7*k observations), and scaling.
- We can further search for better clusters using different evaluation metrics, pre-process differently, and potentially use kmlShape to find groups of characteristics using shape-respecting distance and means.
- In the future, we can use other data streams from the National Weather Service and PurpleAir monitors.

This work was supported by funding from the Undergraduate Research Opportunities Program at the University of Utah awarded to Grace Liu.

THE UNIVERSITY OF UTAH