For the research paper I worked on this spring, I investigated how fire emissions, particulate matter, and climate variables/indices (temperature, precipitation, drought indices) impact mortality rate of chronic lower respiratory diseases.

While cholera is an infectious disease caused by *V. cholerae* bacteria, chronic lower respiratory diseases are not primarily infectious in etiology. They do have infectious determinants, but this is an active area for investigation ([Infections in “Noninfectious” Lung Diseases](https://www.atsjournals.org/doi/full/10.1513/AnnalsATS.201401-041PL)). They are also attributable to other factors, such as smoking and air conditioning. With these caveats in mind, I trained a model using datasets available on the national, county-level scale, then discussed existing studies that address the limitations. These studies are smaller in scale, such as in individual cities. While they lack data on the national scale, they are better suited to study localized effects on mortality.

I was particularly inspired by the paper “[Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables](https://www.mdpi.com/1660-4601/17/24/9378/htm).” The authors used random forest from SciKit Learn to classify cholera outbreaks, with 1-2 month lagged values of climate variables to represent their lagged effects on cholera incubation periods. I also used random forest, except I performed regression rather than classification, and incorporated fire emissions, biosphere fluxes, burned area, particulate matter, climate variables/indices, and their 1-2 month lagged values.

I optimized model performance (maximizing R-squared) using 10-fold cross validation with SciKit Learn's [GridSearchCV](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html) for parameter testing and [RFECV](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html) for feature selection. To further analyze model performance, I also scored a test dataset not used in training.

Data sources:

* [Global Fire Emissions Database](http://www.globalfiredata.org/) (GFED)
  + fire emissions, burned area
  + biosphere fluxes
    - net primary production (NPP), heterotrophic respiration (Rh)
* [Atmospheric Composition Analysis Group](https://sites.wustl.edu/acag/datasets/surface-pm2-5/)
  + particulate matter (PM2.5)
* [NOAA Monthly U.S. Climate Divisional Database](https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00005) (nClimDiv)
  + climate variables/indices (precipitation, temperature, drought indices)
* US Census Bureau ([Census.gov](https://www.census.gov/))
  + population, median income, shapefiles
* [CDC WONDER](https://wonder.cdc.gov/WelcomeT.html); [Underlying Cause of Death, 1999-2019 Request (cdc.gov)](https://wonder.cdc.gov/ucd-icd10.html)
  + County-level death counts from causes described by the International Classification of Diseases (ICD)
  + Extracted death counts for chronic lower respiratory diseases (ICD codes J40-J47)

GitHub repository: [Unusuala1l2e3x4/Research-Spring2021](https://github.com/Unusuala1l2e3x4/Research-Spring2021)