Chronic Respiratory Disease: Risk Modeling Potential and Limitations

Alexander He
hea2@rpi.edu

Thilanka Munasinghe
munast@rpi.edu

IEEE BIG DATA 2021
Virtual Event • 15–18 December

Motivation

- Chronic Respiratory Diseases (CRDs) are among the leading causes of mortality worldwide, with 545 million prevalent cases in 2017
- Symptoms of non-infectious CRDs are often exacerbated by:
  - ambient air pollution
  - changes in temperature and humidity

- Inspired by study where Machine Learning (ML) was successfully used to forecast the risk of Cholera outbreaks in India
  - Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables

Datasets

- Mortality (cause-specific counts of death)
- Population
- Shapefiles (counties and climate divisions)

- Spatiotemporal datasets
  - Fine particulate matter (PM2.5)
  - Carbon emissions, biosphere fluxes, burned area
  - Climate variables, drought indices

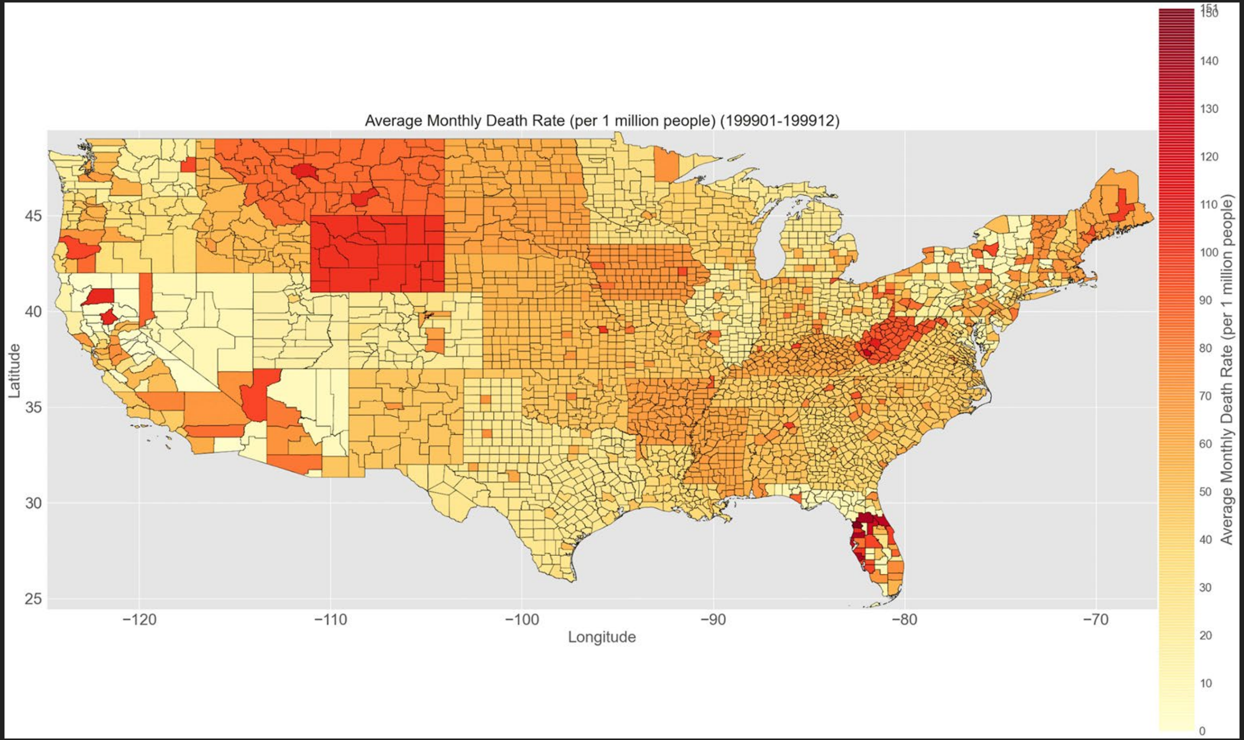
Scope

- Period of interest: 2000 - 2016
- Monthly temporal resolution
- Counties in contiguous U.S.

Mortality - Chronic Lower Respiratory Diseases (CLRDs)

- [CDC WONDER]
  - Underlying Cause of Death

- Includes:
  - asthma, emphysema, bronchiectasis, other COPDs (generally non-infectious)
- Excludes:
  - influenza, pneumonia, other respiratory infections (infectious)



Population

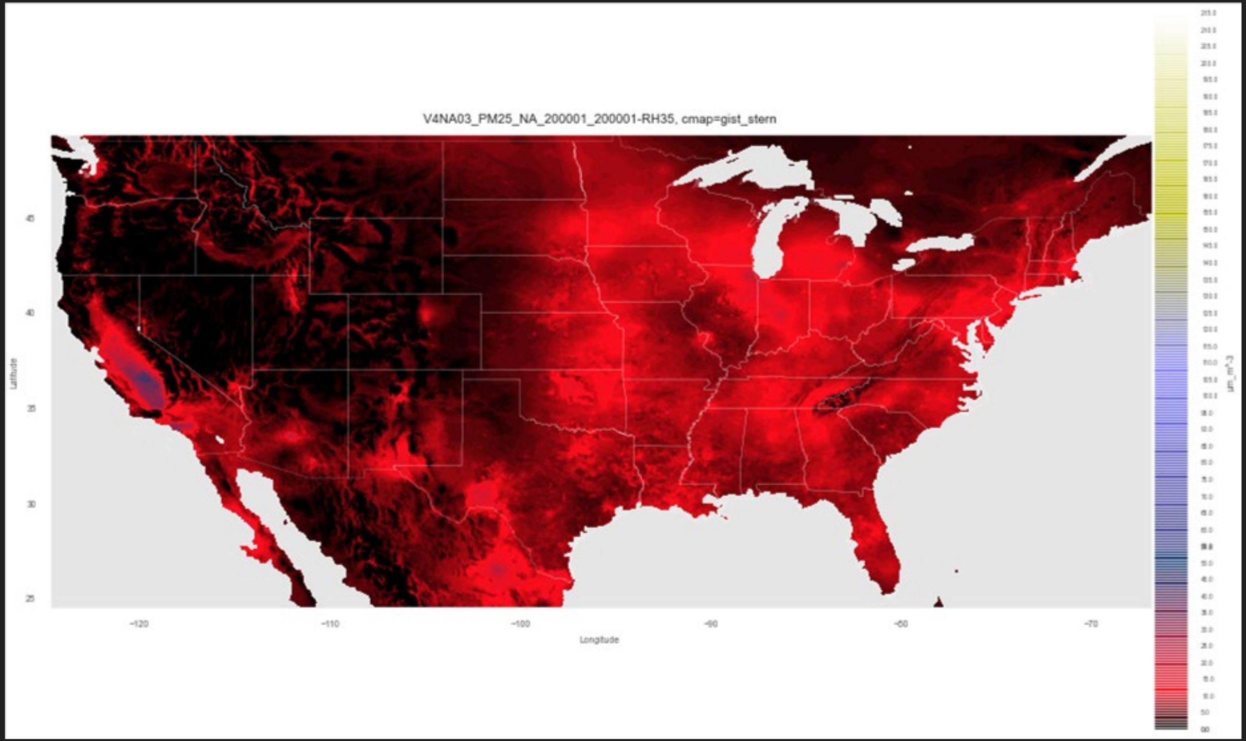
- Datasets - US Census Bureau
- Monthly population totals for each county

- Purpose: calculating mortality rate and population density

Fine particulate matter (PM2.5)

- [Atmospheric Composition Analysis Group]
  - Washington University in St. Louis

- 0.01° × 0.01° grid
- µg m^{-3}



Carbon Emissions, Biosphere Fluxes, Burned Area

- Global Fire Emissions Database (GFED)
- 0.25° × 0.25° grid

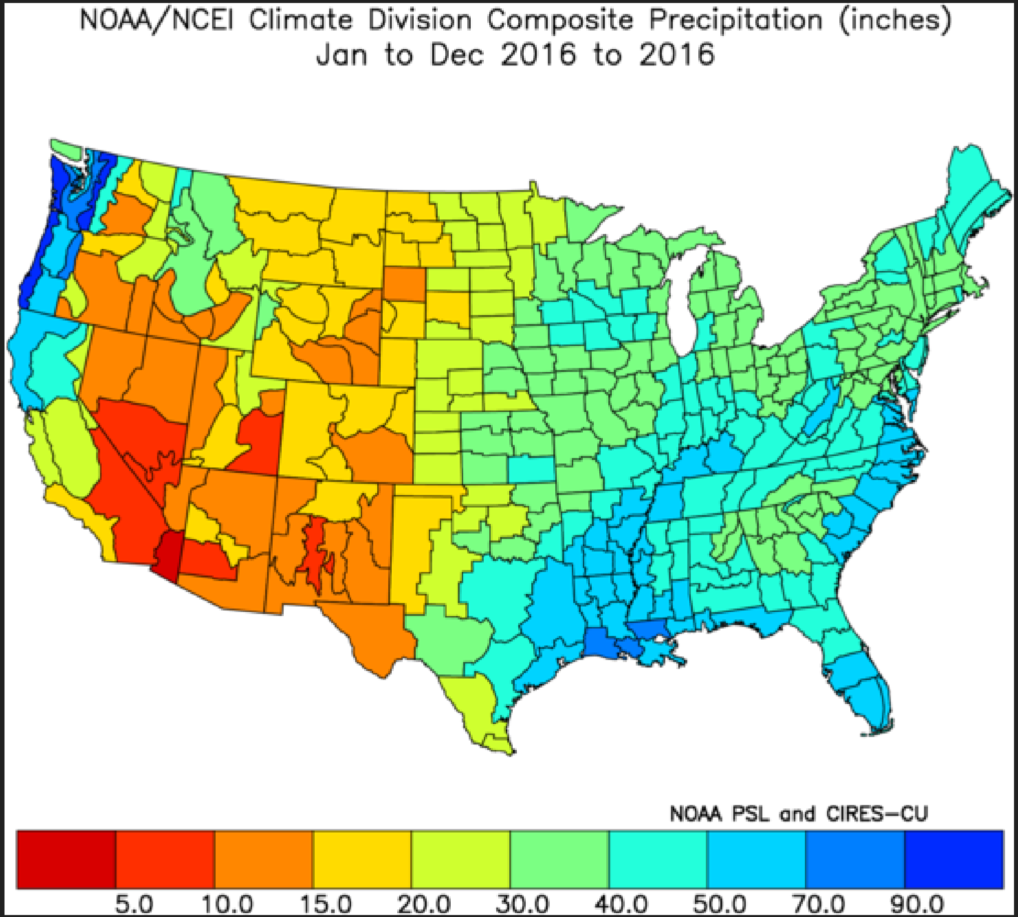
- Carbon emissions g C m^{-2}
- Biosphere Fluxes g C m^{-2}
  - net primary production (NPP)
    - C gained (photosynthesis) minus C released (plant respiration)
  - heterotrophic respiration (R_h)
  - fire emissions (BB)
- Burned area
  - Fraction of each grid cell that burned in each month
  - Actual area - calculated with grid cell area data provided

Climate Variables, Drought Indices

- NOAA Monthly U.S. Climate Divisional Database (NClimDiv)
- By climate divisions

- Climate variables
  - Temperature
  - Precipitation
- Drought indices; negative = dry spells, positive = wet spells
  - Palmer Drought Severity Index (PDSI); -6 to +6
    - balance between moisture supply and demand.
  - Standardized Precipitation Index (SPI, SP01 for monthly); -3 to +3
    - 0 = median of precipitation for particular location

Total Precipitation, 2016



NOAA/NCEI U.S. Climate Division Data Plots: NOAA Physical Sciences Laboratory

Shapefiles - Counties, Climate Divisions

- Boundaries - collections of points; polygons
- Counties - [Cartographic Boundary Files - US Census Bureau]

- Metadata
  - Location codes for county and state
  - Land + water area of each county

- Purpose:
  - Determining the grid cells in each county (or climate division)
  - Calculating county population density

Data Preparation

- Convert to identical 0.01° × 0.01° grid beforehand
- Aggregate all spatiotemporal datasets by county and month
    - Adjust by total area of grid cells in each county
- Include 1- and 2-month lags for spatiotemporal variables
    - Fine particulate matter (PM2.5)
    - Carbon emissions, biosphere fluxes, burned area
    - Climate variables, drought indices

- Update county boundaries and designations
    - Changes to Counties and County Equivalent Entities: 1970-Present - US Census Bureau

Methods

- 70:30 train-test split

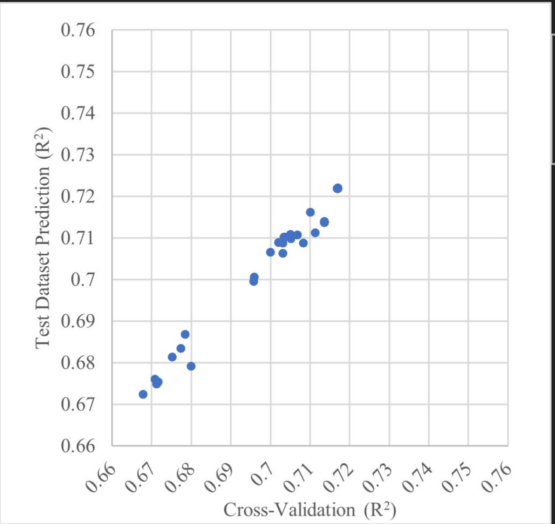
- Random forest regression (scikit-learn)
  - 10-fold cross validation
  - Hyperparameter tuning to optimize model
  - Feature selection - recursive feature elimination
  - Optimize R-squared

- Collinearity analysis with Spearman rank correlation (SciPy)
  - Improves discussion of variables' potential contributions to the model

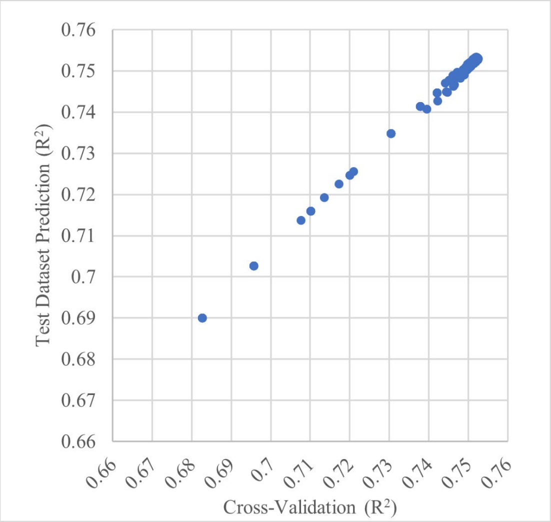
Results

- R-squared
  - 0.7526 - cross validation
  - 0.7528 - test dataset prediction

- Similar trends between cross-validation and test dataset prediction
  - Suggests model generalizes well for unseen data



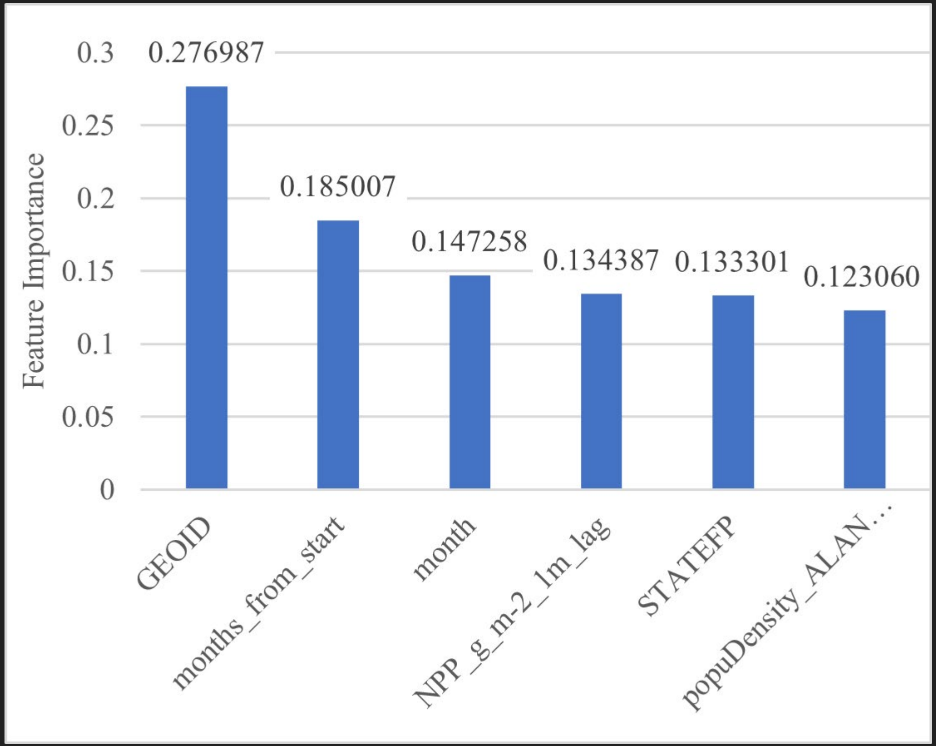
During RFECV iterations



During hyperparameter tuning (after RFECV)

Results - Selected Features

1. GEOID - county encoder
2. Months from start of period (since January, 2000)
3. Month of the year
4. Net primary production (NPP), lagged by 1 month
5. STATEFP - state encoder
6. Population density, adjusted by land area



Data Limitations

- Using mortality as target variable
  - Mortality only captures most extreme cases of disease exacerbation
  - Limited data of Emergency Room Visits (ERVs) by county
  - ERVs by country - commonly used to measure disease exacerbation
    - Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits

- Mortality data - suppression constraints
  - Data points with less than 10 deaths are unavailable
  - Estimated based on state total

- Monthly, county-level datasets not available for:
  - Humidity
    - Used precipitation, drought indices instead
  - Ground-level ozone
    - estimated asthma ERVs in 2015:
      - Ozone: 9–23 million
      - PM2.5: 5–10 million

Other Limitations

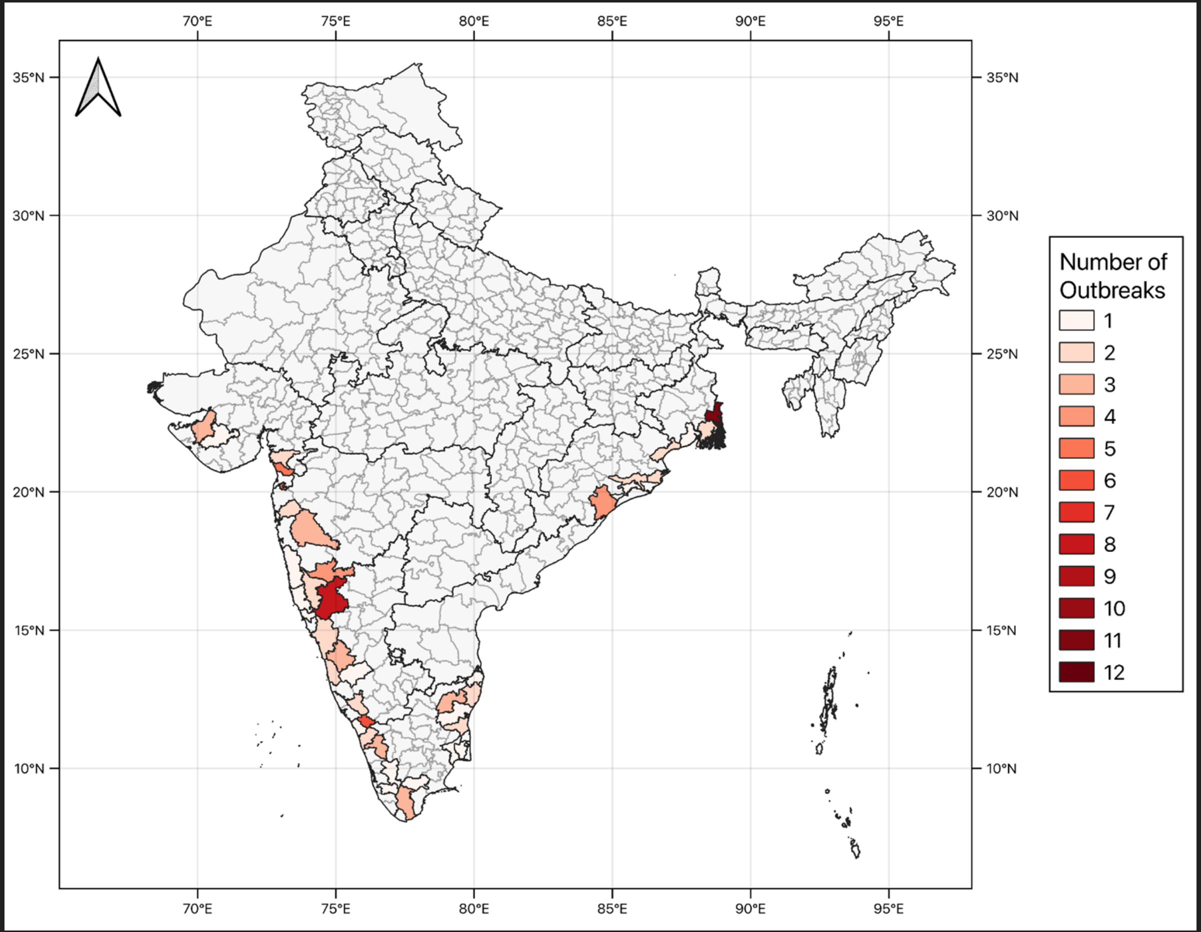
- Coarse resolution - monthly, county-level datasets cannot represent phenomena present in finer spatiotemporal resolutions (i.e. cities, days)
    - Air pollution and emergency department visits for respiratory diseases: A multi-city case crossover study
        - ERVs for respiratory diseases increase after days with higher concentrations of pollutants
        - Lagging effect on ERVs lasts only several days

- Excluded factors
    - tobacco smoking (human behavior)
    - airborne allergens
    - Many “non-infectious” lung diseases are influenced by infectious microorganisms
        - active area of investigation

Comparison - Cholera Outbreaks in India

- Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables
- Random forest classification (scikit-learn)
- Essential Climate Variables (ECVs)
  - Chlorophyll-a, Land/Sea Surface Temperature, Soil Moisture, Total Precipitation, etc
- *Vibrio cholerae*
  - Infectious; prevalent in coastal areas
- Strong spatiotemporal relationships between ECVs and distribution of *V. cholerae* bacteria
  - Allows accurate predictions of cholera outbreaks
  - Sensitivity = 0.895; correctly identified 89.5% of outbreaks
  - ROC = 0.984

Number of cholera outbreaks reported; 40 coastal districts; 2010 - 2018



Main conclusions

- Significant potential for modeling CRD risk (R-squared ≈ 0.75)
- Limitations
  - data limitations
  - phenomena only found on finer spatiotemporal scales (i.e. cities, days)
  - primarily non-infectious nature of CRDs
- Caveats of ML
  - can naively test for features that improve model performance
  - mechanisms behind their relationships remain unclear

- Accessible data of climate variables may improve estimates of ERVs attributable to specific pollutants
  - Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits

Next Steps

- Use better methods for hyperparameter tuning
  - Gradient-based rather than grid search
- Address asymmetry of importance
  - imbalanced regression
- Use different ML techniques
  - neural networks

- Use datasets from alternative sources
  - Datasets may be developed differently
  - New datasets may appear in the future