Using Machine Learning and Statistics to Understand Correlations in COVID-19 Mortality

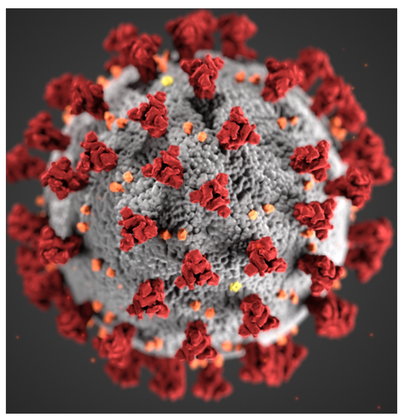
Advanced Science and Engineering Program
Ms. Boylan, Mr. Renauld, Mr. Wardop
Rishi Basu (rishi.basu@sps.edu)

Abstract

The novel coronavirus, COVID-19, caused a pandemic that captured the world's attention. In the United States as of writing, it has taken over 190,000 lives. It has also shown parallels with other events of 2020, as COVID-19 mortality has disproportionately affected certain higher-risk and underserved populations, provoking necessary discussions on social issues in the US. Using data to identify correlating variables with COVID-19 mortality is a global epidemiological concern.

COVID-19 also directly shaped my summer plans. Since Harvard Medical School closed down its research labs to repurpose them for clinical work, I was unable to research how cells measure mechanical pressure at a systems biology lab with Professor Markus Basan.

Focusing on the subject itself, I built upon research at the Harvard T.H. Chan School of Public Health (HSPH) investigating whether pollution increases the risk of COVID-19 deaths in the United States (Wu et. al). I spent the summer following HSPH's studies and searching for potential environmental and socioeconomic confounders of COVID-19 mortality.



An illustration of the novel coronavirus

Data

The data I used was collected from various sources. Including:

- Johns Hopkins University Center for Systems Science and Engineering
- US Census Bureau
- CDC Behavioral Risk Factor Surveillance System
- Carnegie Mellon University Delphi Epidemiological Group

HSPH's original results were from **April 4th, 2020** when there were **~10,000 COVID-19 deaths** in the United States, and **77.8%** of US counties had zero deaths.

Updated data for **August 19th, 2020** shows **~175,000 COVID-19 deaths** with only **24.5%** of counties remaining death-free.

Choropleth maps: Figures 1-3

Fig. 1 is a choropleth map of pollution levels measured by average particulate matter concentrations by county.

There is some noticeable correlation between Fig. 1 and Fig. 2 & 3, suggesting that higher exposure to particulate matter and pollution may increase risk of COVID-19 mortality.

Note: Fig. 2's data is in natural log due to outliers and greater dispersion.

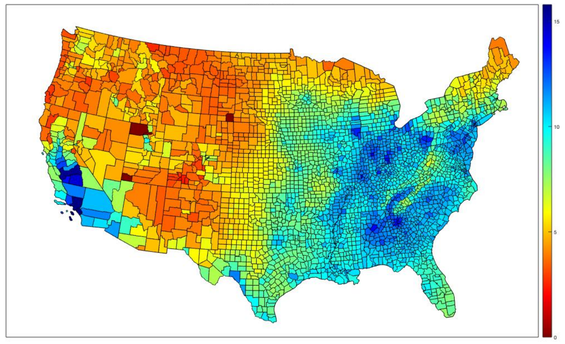


Fig. 1- County Level 17-year Long-Term Average of PM2.5 Concentrations in the US in micrograms/m^3

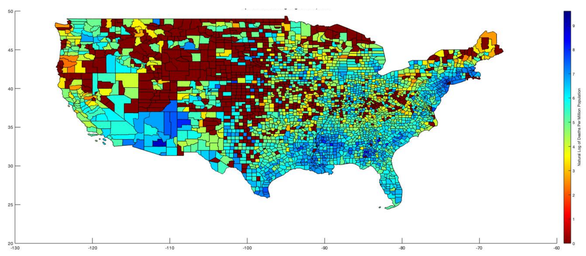


Fig. 2- County Level Natural Log of Number of COVID-19 Deaths Per One Million Population in the US Up To August 19th, 2020

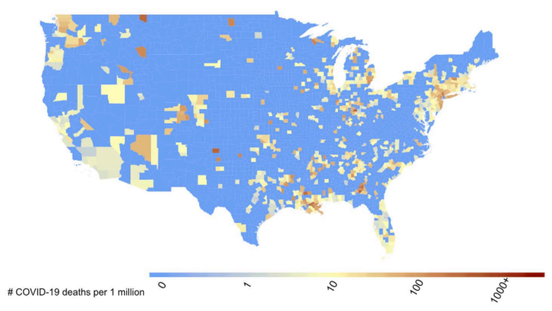
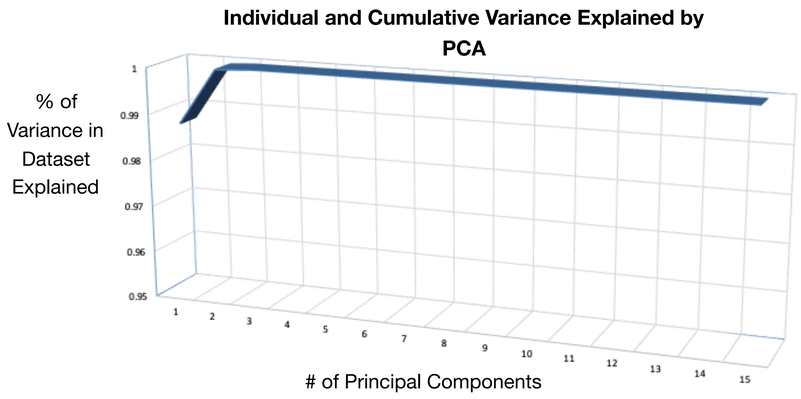


Fig. 3- County Level Number of COVID-19 Deaths Per One Million Population in the US Up To April 4th, 2020 (Wu et. al)

Principal Component Analysis (PCA)



1. **Dataset with 15 county-level features** (COVID-19 mortality rate, pollution, weather, poverty, household value, percent owner occupied, population density, race, percent elderly).
2. **PCA– Unsupervised learning algorithm** for **dimensionality reduction** and compacting data
   - Each datapoint begins by being represented on 15 axes (one per feature).
   - PCA **rotates the axes** so that they become **linear combinations** of the original features to **minimize distance** between **new axes**, or, **principal components**, and **data points.**
   - The remapped axes are called **eigenvectors** and are listed in decreasing importance. The **eigenvalues** explain the variance of the data. Dividing each eigenvector's eigenvalue by the sum of all eigenvalues reveals the proportion of the variance that each eigenvector represents.
   - With my dataset, **the first eigenvalue divided by the sum of all eigenvalues was 98.7%** (see chart). This means that the features are highly related. If they were completely independent of each other, the slope of the line would be 45 degrees from 0% to 100% variance explained, and the first principal component would only describe 100/15=6.67% of the variance.
   - This means that any feature, including COVID-19 mortality, could be accurately predicted by a linear combination of the other features. Implies correlation between COVID-19 mortality and mentioned environmental & socioeconomic features.

Works Cited

Wu, X., R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici (2020): "Exposure to air pollution and COVID-19 mortality in the United States," medRxiv.

HARVARD MEDICAL SCHOOL

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH