Can Certain Health Problems Be Predicted by Environmental Factors?

Introduction

There is a push in the U.S. to address environmental injustices. Overburdened communities are defined as "minority, low-income, tribal, or indigenous populations or geographic locations in the United States that potentially experience disproportionate environmental harms and risks."[i] To determine which are the most pressing environmental problems, the harm from particular environmental problems should be quantified in terms of health effects.

The goal of this research was to predict specific negative health outcomes: asthma prevalence, cancer, chronic kidney disease, chronic obstructive pulmonary disease (COPD), coronary heart disease (CHD), and stroke from a variety of environmental factors. This prediction would allow these outcomes to be linked to specific environmental factors in overburdened communities.

The dataset with the eleven environmental indicators came from the EJScreen [environmental justice screen] 2020 of the U.S. Environmental Protection Agency.[ii] The abbreviations for these indicators are defined in the appendix. The data on the health outcomes were obtained from U.S. Centers for Disease Controls and Prevention's Local Data for Better Health 2022,[iii] which is based on 2020 information.

Data Wrangling, EDA, and Pre-processing

Since the health data was collected at the census tract level and the environmental data was collected at the census block level, data wrangling was needed in order to combine the two datasets. The environmental data was converted to census tracts by combining the blocks within each tract by population-weighted averaging.

No correlation was found between the health outcomes and any one of the environmental risk factors. For example, Figure 1 shows a heat map of COPD. Only the lead paint indicator (PRE1960PCT) is greater than 0.10 positively correlated with the prevalence of COPD; however, this correlation is very, small. Diesel fuel particulates (DSLPM), proximity to and volume of traffic (PTRAF), and proximity to treatment storage and disposal facilities (PTSDF) are all less

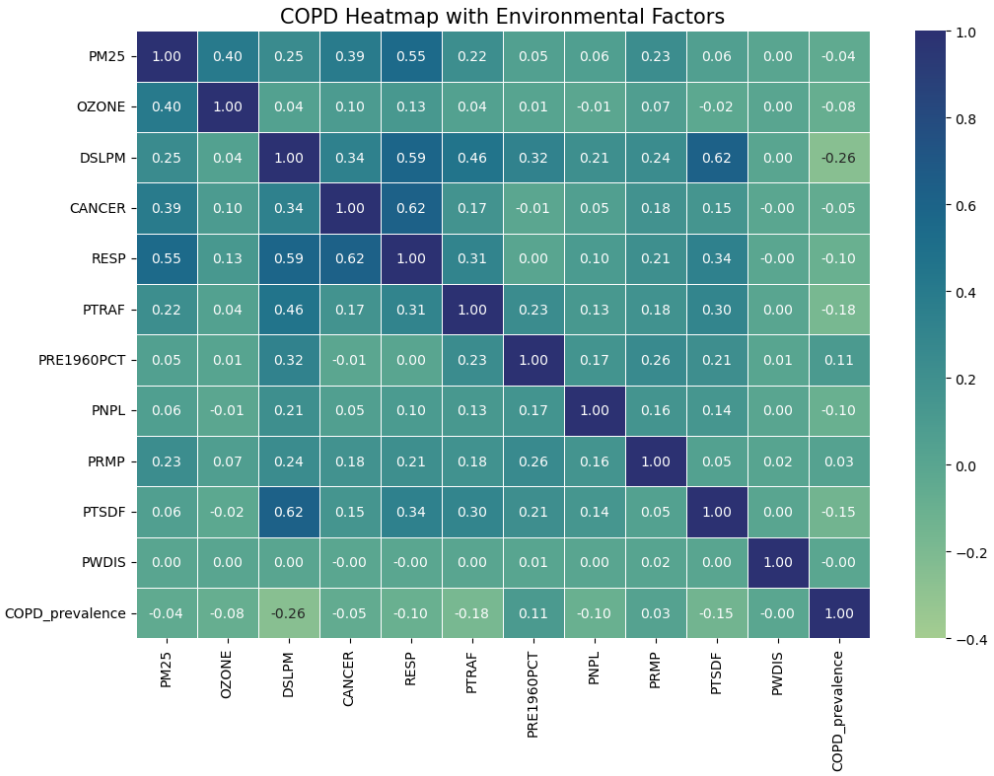

Figure 1. A heat map of COPD with environmental factors.

than -0.10 negatively correlated to COPD, with DSLPM having the most negative correlation of -0.26. None of these correlations is very strong and therefore modeling was done with all (or almost all) of the environmental features).

Given how skewed the data was, the best way found to normalize it was to take the log. Since PM25 and OZONE distributions were already reasonably normal to begin with, they were not transformed. The environmental data before normalization is shown in Figure 2, and after normalization in Figure 3. Since PWDIS, the indicator for major direct dischargers to water, could not be normalized by any technique, this indicator was dropped from the models that require normalized data.

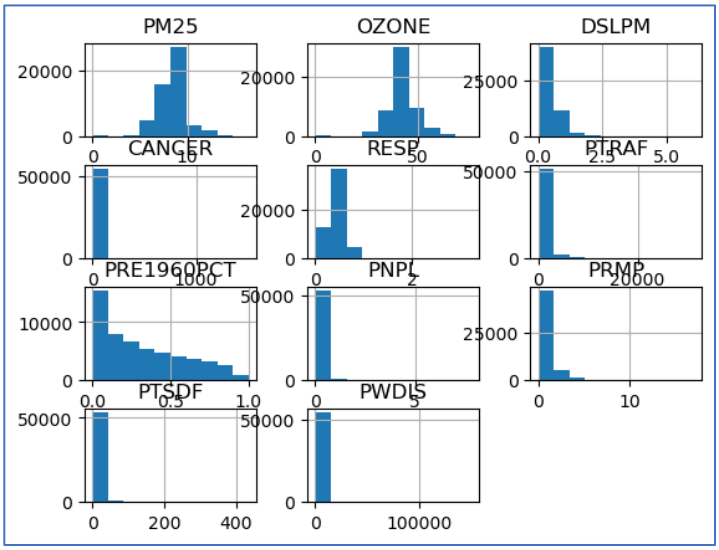

Figure 2. Histograms showing the distribution of data for each environmental factor before normalization.

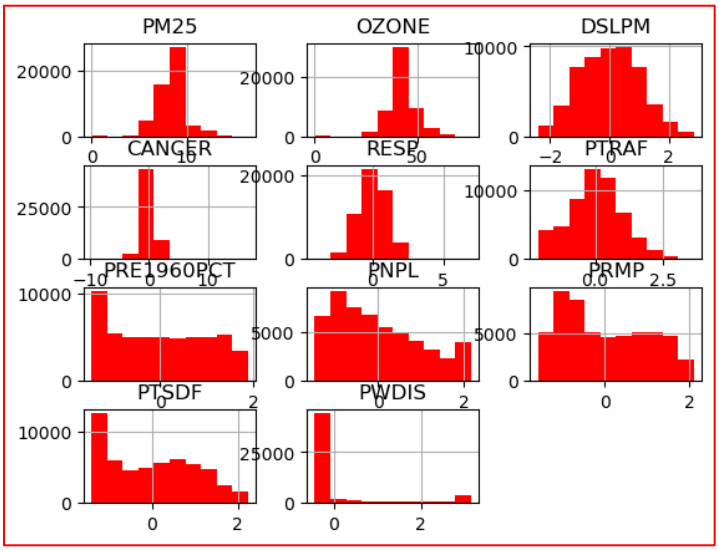

Figure 3. Histograms showing the distribution of data for each environmental factor after normalization (except PM25 and. OZONE have not been normalized.

Modeling

The first model used was the simplest: multiple ordinary least squares (OLS). Since this was not a good predictor for any of the health outcomes, the following ensemble decision tree regressors were tried next: Random Forest, AdaBoost, GradientBoost, XGBoost, and LightGBM. A support vector machine (SVM) regressor was also used. Since only the SVM required the normalized data, the rest of the model were run with the unnormalized data. Some hyperparameters were tuned via randomized search and some via Bayesian search. Figure 4 lists the best model for each health outcome, and that model's the R^2 value, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), along with any tuned hyperparameters. Figure 5 is a scatter plot of the predicted values of COPD versus the actual values using the model shown in Figure 4.

	Best Model	Type of Tuning	R^2	MAE	MSE	RMSE	Tuned Parameter values
COPD	Random Forest	Bayesian	0.568	1.26	2.98	1.73	max_depth=36, n_estimators=600
Asthma	XGBoost	Bayesian	0.554	0.758	1.07	1.03	min_child_weight=44, max_depth=
CHD	Random Forest	Randomized	0.513	0.970	1.90	1.38	max_depth=27, n_estimators=350
Stroke	Random Forest	none	0.506	0.585	0.742	0.861	
Kidney	Random Forest	none	0.481	0.515	0.516	0.719	
Cancer	Random Forest	Randomized	0.455	0.941	1.84	1.36	max_depth=24, n_estimators=134

Figure 4. Modeling data.

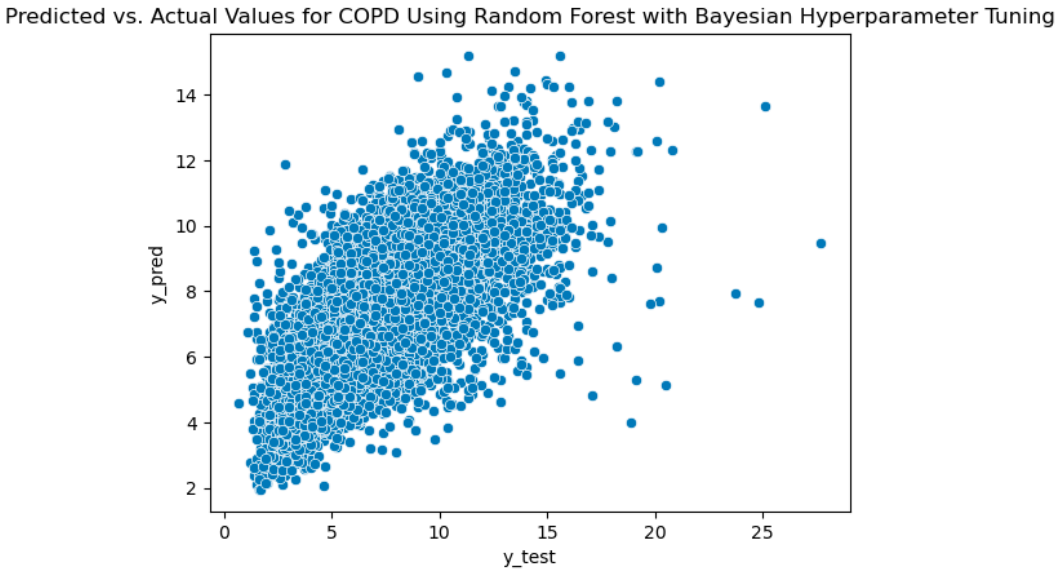

Figure 5. Scatter plot of the predicted values of COPD versus the actual values from the model shown in Figure 4.

Analysis

None of the models used in this research were good predictors of any of the health outcomes. The best R^2 values ranged from 0.455 for cancer to 0.568 for COPD and came from Random Forest models except for asthma, for which XGBoost was the best model. Figure 5 shows this problem clearly: the maximum predicted values are less than 16 while the actual values go up to 25.

The most likely reason that there was low correlation between the environmental features and the health outcomes most likely comes from the health outcomes data used. These data were collected by asking people, "Have you ever been told by a doctor, nurse, or other health professional that you have [insert health outcome here]?" Self-reported data is not always reliable, and the CDC itself states about the asthma data: "Physician-diagnosed asthma is self-reported in the Behavioral Risk Factor Surveillance System and was not confirmed by a health-care provider or objective monitoring. This survey-based indicator requires a doctor diagnosis of asthma, which may not include all persons with asthma."[iv]

Another reason is that while certain environmental factors are known to increase one's chances for certain diseases, there is a time factor. Comparing health and environmental data from the same year and geographical area intuitively makes sense, but it does not consider the environment in which the people were over time prior to getting the diseases.

Future Work

To get information about the types of diseases caused by these environmental factors, the next steps would be to obtain objective health data, to include only people who have been living in the same census tract for 10 years, and to use more than one year of environmental data.

Appendix

Definitions of Environmental Risk Factors:

PM25 = Particulate matter 2.5 level in air

OZONE = Ozone level in air

DSLPM = Diesel particulate matter level in air

CANCER = Air toxics cancer risk

RESP = Air toxics respiratory hazard index

PTRAF = Traffic proximity and volume

PRE1960PCT = % pre-1960 housing (lead paint indicator)

PNPL = Proximity to National Priorities List (NPL) [superfund] sites

PRMP = Proximity to Risk Management Plan (RMP) facilities, facilities that use extremely hazardous substances

PTSDF = Proximity to Treatment Storage and Disposal facilities (TSDF)

PWDIS = Indicator for major direct dischargers to water

[i] Environmental Protection Agency, https://www.epa.gov/environmentaljustice/ej-2020-glossary.
[ii] Environmental Protection Agency, https://gaftp.epa.gov/EJScreen/2020.
[iii] "PLACES: Local Data for Better Health, Census Tract Data 2022 release", Centers for Disease Control, https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/cwsq-ngmh.
[iv] "Health Outcomes Measure Definitions" in "PLACES: Local Data for Better Health, Census Tract Data 2022 release", https://www.cdc.gov/places/measure-definitions/health-outcomes/index.html#asthma.