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.”[[1]](#endnote-1) 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.[[2]](#endnote-2) 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,[[3]](#endnote-3) 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 positively correlated with the

A graph of copd heatmap with environmental factors

Description automatically generated with low confidence

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

prevalence of COPD; however, this correlation is very, small. Given how skewed the data was, the best way found to

normalize it was to take the log. Since PM25 and OZONE 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.

A picture containing text, diagram, plan

Description automatically generatedA picture containing text, diagram, plan, line

Description automatically generated

**Figure 2**. Histograms showing the distribution of data for **Figure 3**. Histograms showing the distribution of data for

each environmental factor before normalization. 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 R2 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.

A screenshot of a computer

Description automatically generated

**Figure 4**. Modeling data.

A picture containing text, screenshot, diagram, line

Description automatically generated

**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 R2 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.”[[4]](#endnote-4)

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

1. Environmental Protection Agency, <https://www.epa.gov/environmentaljustice/ej-2020-glossary>. [↑](#endnote-ref-1)
2. Environmental Protection Agency, <https://gaftp.epa.gov/EJScreen/2020>. [↑](#endnote-ref-2)
3. “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>. [↑](#endnote-ref-3)
4. “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>. [↑](#endnote-ref-4)