Where Can Clean Technology Help? Machine Learning to Identify Environmentally At-Risk Communities in the United States

Shiran Shen
srshen@stanford.edu
Blane Wilson
bctwilson@stanford.edu
Stanford University

Abstract

Inspired by CalEnviroScreen, an environmental health assessment tool used to identify environmentally at-risk communities in California, we calculate pollution burden scores at the census tract level for the entire contiguous United States. Pollution burden is a composite score that encompasses 12 environmental (air, water, waste) indicators. We combine actual pollution burden indicator data with predicted statistics using machine learning. We create a novel National (Lower 48) Pollution Burden Map using ArcGIS.

1. Introduction

California arguably has the most comprehensive environmental health data collection, which significantly aids its efforts to identify communities most affected by a variety of pollution sources, home to residents most vulnerable to the adverse effects of pollution. The Office of Environmental Health Hazard Assessment (OEHHA) has developed an environmental health assessment tool, called "CalEnviroScreen," for all census tracts in California. It significantly aids California's efforts to identify communities that can benefit most from clean technology. While the CalEnviroScreen final score is computed with the product of the pollution burden score and the vulnerable population score, we focus on the pollution burden score (i.e., exposure to pollutants, environmental effects). Our goal is to develop a pollution burden scoring system at the census tract level for the entire contiguous U.S.

Motivated by existing evidence suggesting that underprivileged groups tend to live in more polluted communities, we use relevant sociodemographic indicators as predictors for pollution burden indicators whose data are unavailable for other states. We apply multiple regression models and choose the one with the least cross-validation (CV) test error on California data to make predictions for missing pollution indicator data for other states. Following the methodology in CalEnviroScreen, we calculate and map pollution burden scores for census tracts in every continental US state.

2. Related Works

Many existing works point to the startling fact that pollution inequality is more severe than income inequality in the U.S. Socioeconomic and racial/ethnic status has been frequently linked to exposures to environmental risks, such as proximity to hazardous waste sites, vehicle traffic, and polluting industries [1, 8, 6, 2, 4]. Nationally, inequality for NO_2 concentration is greater than inequality for income [3]. These socioeconomically underprivileged communities are often made more susceptible to air pollution due to poor access to health care [7]. We leverage this widely documented correlation between relevant sociodemographic indicators and pollution outcomes to make predictions for missing pollution indicator data.

3. Data

Data on pollution burden indicators at the census tract level of California come from CalEnviroScreen 3.0, released in September 2016. It comes on a 0-10 scale and is multiplied by 10 to obtain a scale of 0-100 in our training dataset. As shown in Figure 1, there are in total, twelve pollution burden indicators that encompass both exposure and environmental effects.

For other states, we collected their pollution burden indicator data at the census block group level from the Environmental Protection Agency (EPA) and the Department of Transportation (DOT). We were able to collect data for all pollution burden indicators to match those in CalEnviroScreen except for pesticide use, drinking water contaminants, and solid waste sites and facilities (highlighted in yellow in Figure 1); this is where we use machine learning to make predictions for these missing indicator data. We aggregated census block-level data to obtain census tract-level data to match the administrative level at which CalEnviroScreen scores are calculated. Finally, we calculated the state percentiles for each indicator.

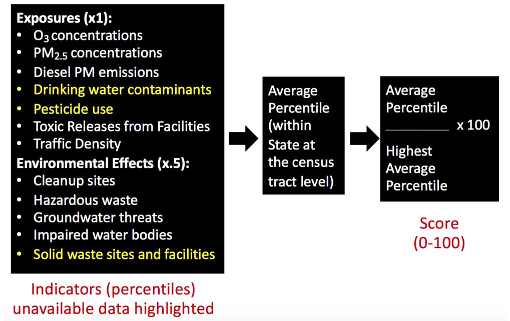

Exposures (x1):
O$_3$ concentrations
PM$_{2.5}$ concentrations
Diesel PM emissions
Drinking water contaminants
Pesticide use
Toxic Releases from Facilities
Traffic Density
Environmental Effects (x.5):
Cleanup sites
Hazardous waste
Groundwater threats
Impaired water bodies
Solid waste sites and facilities
Indicators (percentiles) unavailable data highlighted
Average Percentile (within State at the census tract level)
Average Percentile
x 100
Highest Average Percentile
Score (0-100)

Figure 1. Environmental indicators and pollution burden score

As mentioned in related works, given a high correlation between sociodemographic variables and pollution outcomes, we collect data on relevant variables, including poverty, known poverty status, minority, less than high school education, under 5 years old, over 64 years old, household linguistic isolation, and population density, at the census block level from the 2010-2014 American Community Survey (ACS), a national demographics survey administered by the U.S. Census Bureau. Specifically, they are proportion of population who are in households with household income no more than twice the federal "poverty level;" proportion of the population whose poverty (or lack thereof) is known; proportion of individuals whose race status is something other than non-Hispanic white alone; proportion of the population that are over 25 years old with less than high school degree; percent of people who are under 5 years old; percent of people who are over the age of 64; the proportion of households where all members speak English less than "very well;" and population size divided by land area. As with pollution burden indicator data, we aggregated these census block group-level data to obtain census tract-level data. We then calculated state-level percentiles for these indicators except known poverty status, which is in percentage. Known poverty status is an indicator of the coverage of the poverty survey, so it is more appropriate to use its percentage rather than percentile form.

4. Methods

Here we aim to predict percentile values for indicators with unavailable data in other states. The sociodemographic indicators detailed in Section 3 represent our predictors. We compare four different models with the traditional, baseline OLS model that incorporates all features: ridge, lasso, best subset selection model, and best subset selection model with interaction terms between selected features. Figure 2 demonstrates our workflow. Individual steps are detailed below.

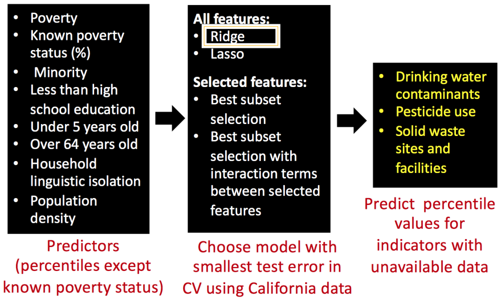

Poverty
Known poverty status (%)
Minority
Less than high school education
Under 5 years old
Over 64 years old
Household linguistic isolation
Population density
Predictors (percentiles except known poverty status)
All features:
Ridge
Lasso
Selected features:
Best subset selection
Best subset selection with interaction terms between selected features
Choose model with smallest test error in CV using California data
Drinking water contaminants
Pesticide use
Solid waste sites and facilities
Predict percentile values for indicators with unavailable data

Figure 2. Environmental indicators and pollution burden score

4.1. Create training and test samples for California data

As our initial step, we split our California dataset in half randomly to create the training and the test samples for calculation of CV error for different models in step 4.3.

4.2. Train algorithms

We train five different algorithms on the California training sample: full-feature OLS, ridge, lasso, best subset-selected features OLS, and best subset-selected features with interaction terms OLS. In the full-feature OLS model, the goal is to minimize the cost function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2$$

Similar to OLS, ridge regression estimates coefficients by minimizing a slightly different quantity. The $\hat{\beta}^R$ values minimize:

$$\sum_{i=1}^{m}(y^{(i)} - \beta_0 - \sum_{j=1}^{n}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{n}\beta_j^2 = J(\theta) + \lambda\sum_{j=1}^{n}\beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter that is determined separately. It trades off two different criteria [5]. On the one hand, the first term seeks to estimate coefficients that fit the data well by making the cost small, just like in OLS. On the other hand, the second term, $\lambda\sum_j \beta_j^2$, is called the "shrinkage penalty;" it is small when the coefficients are close to zero, thereby shrinking the coefficient estimates towards zero. The tuning parameter, λ, controls the relative impact of these two terms on the coefficient estimates. Unlike OLS, which produces only one set of coefficient estimates, ridge produces a distinctive set of coefficients,

$\hat{\beta}_{\lambda}^{R}$, for each value of λ. Choosing the right value for λ is critical. Ridge regression's advantage over OLS is rooted in the bias-variance trade-off. The increase in λ is accompanied by decreased variance and increased bias.

By contrast, lasso may penalize and force some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Lasso coefficients, $\hat{\beta}\,_{\lambda}^{L}$, minimize:

$$\sum_{i=1}^{m}(y^{(i)} - \beta_0 - \sum_{j=1}^{n}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{n}|\beta_j| = J(\theta) + \lambda\sum_{j=1}^{n}|\beta_j|.$$

It differs from ridge regression penalty in that β_j^2 is replaced by $|\beta_j|$ [5]. As with ridge regression, choosing a good value for λ is crucial. In both ridge and lasso regressions, we choose the best λ values using the glmnet R package.

In best subset selection, we fit a separate least squares regression for each possible combination of the eight predictors. To start off, we have exactly one predictor in the model, which gives us eight models. Then we have exactly two predictors in the model, which gives us $\frac{8*(8-1)}{2}$ = 28 models. And so on so forth. There are 2^8 = 256 models in total. We then look at all the resulting models to identify the best one. To state the algorithm more formally, three major steps are involved [5]:

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For k = 1, 2, . . ., p :
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it M_k. Here best is defined as having the smallest J(θ), or equivalently largest R^2.
3. Select a single best model from among M_0, . . . , M_p using cross validated prediction error, Cp (AIC), BIC, or adjusted R^2.

Step 2 is reducing the problem from choosing 256 models to choosing 8+1=9 models. To select a single best model among these 9 models, we want to avoid overfitting by using cross-validated prediction errors, adjusted R^2, Cp, and BIC.

The fifth model we employ simply involves the selected features from best subset selection, plus their interaction terms.

4.3. Generate cross-validation error

We apply the five algorithms on California test sample to generate CV errors for three pollution burden indicators: pesticide use, drinking water contaminants, and solid waste sites and facilities.

4.4. Choose the most appropriate model

Ridge regression model generates the lowest CV error for drinking water contaminants as well as for solid waste sites and facilities. In the case of pesticide use, the difference in CV errors for ridge and lasso is less than one, which is almost negligible. Given that California is more heterogenous in terms of demographics than most other states in the United States, we choose ridge over lasso because the former places less weight on minority than the latter. Therefore, for predicting all three pollution burden indicators, we choose the ridge regression model.

4.5. Predict and adjust percentile values for indicators with unavailable data

Lastly, we predict percentile values for the three pollution burden indicators with missing data using the ridge algorithm we trained in the previous step. Maintaining the relative percentile ranking among census tracts in any given state, we adjust the percentile values so that they are evenly distributed between 0 and 100.

5. Results

We create a national (lower 48) pollution burden map using the platform provided by ArcGIS Online. We map the pollution burden status and sociodemographic characteristics for all census tracts in states with fewer than 1,500 census tracts. Due to data upload capacity of the platform, for states with more than 1,500 census tracts (i.e., AZ, CA, FL, GA, IL, IN, MI, NJ, NY, NC, OH, PA, TX, VA), our map highlights areas with predicted pollution burden scores higher than one standard deviation above the mean (i.e., 15.9 % most polluted tracts). We enable the "Layers" functionality for users to choose which state to display or hide. Access our interactive map here: `http://arcg.is/2gUtEIu`. Below, we showcase a few examples of our map findings.

5.1. Seattle Metropolitan Area, Washington

As shown in Figure 3, pollution burden in the state of Washington is mostly concentrated in the Seattle Metropolitan Area and, to a lesser extent, in Spokane.

Zooming in, we see in Figure 4 that Seattle has the most polluted census tracts in the Metropolitan Area. Zooming in further, we can see that census tract 53033009300 has a pollution burden score as high as 93.30. In this census tract, the percentage of racial minorities, people over 25 with less than high school education, and linguistically isolated households are also among the highest in the state of Washington.

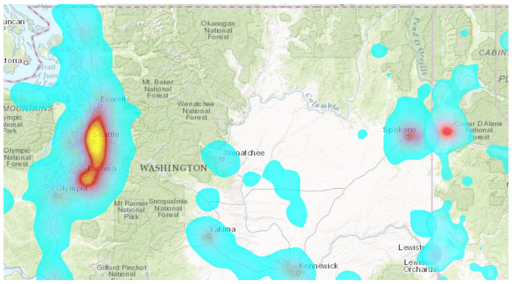


Figure 3. Pollution Burden in the State of Washington

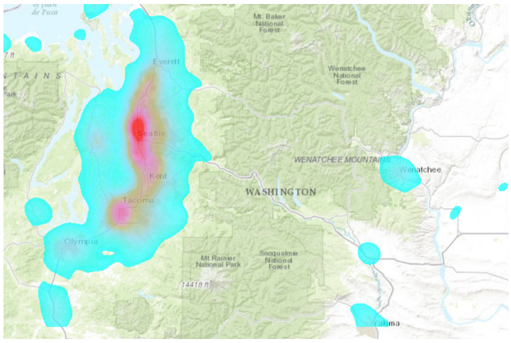
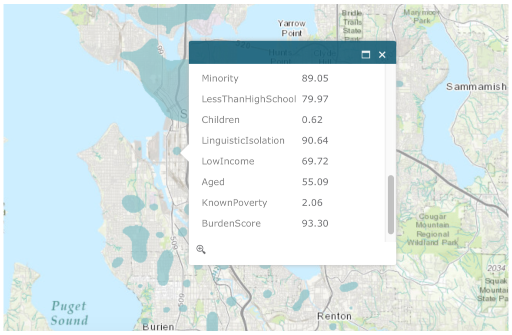


Figure 4. Pollution Burden in the Seattle Metropolitan Area



Minority	89.05
LessThanHighSchool	79.97
Children	0.62
LinguisticIsolation	90.64
LowIncome	69.72
Aged	55.09
KnownPoverty	2.06
BurdenScore	93.30

Figure 5. High Pollution Burden Tract in Seattle

5.2. Dallas, Texas

In the state of Texas, Dallas and Houston are among the most polluted Metropolitan Areas (Figure 6). Unlike its neighbors, Austin has a comparably lower pollution burden level and none of its census tracts are among the 15.9 % most polluted ones in Texas.

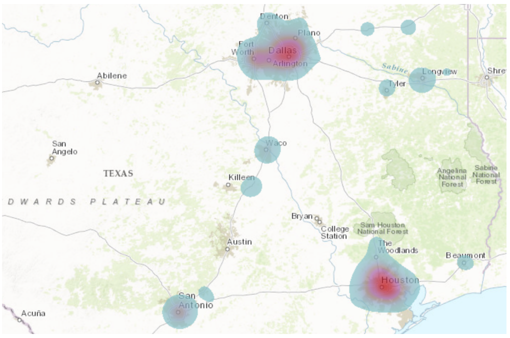
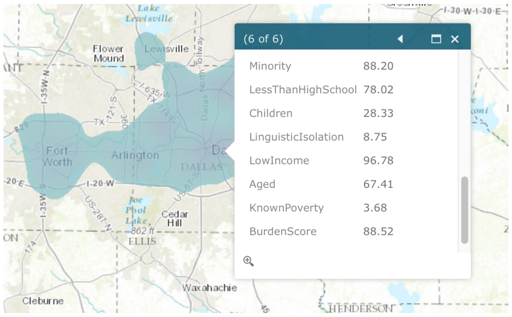


Figure 6. Pollution Burden in the State of Texas

Zooming in at Dallas and clicking on one of the hot spots, we can see that census tract 48113020300 has a pollution burden score of 88.52 (Figure 7). This census tract is also home to widely low income (at 96.78th state percentile) and less educated (78.02th percentile) minorities (88.20th percentile).



(6 of 6)

Minority	88.20
LessThanHighSchool	78.02
Children	28.33
LinguisticIsolation	8.75
LowIncome	96.78
Aged	67.41
KnownPoverty	3.68
BurdenScore	88.52

Figure 7. High Pollution Burden Tract in Dallas

Our interactive map also includes a search box where users can input their address or zip code to view the pollution burden status and sociodemographic composition of their neighborhood.

6. Future Steps

We want to identify ways to validate predicted indicator percentile values. The possibility of this depends on data availability and accessibility. Furthermore, there are a few census tracts with partially missing data, so we want to develop a multiple imputation model for missing data in these census tracts. While we recognize that the validity of our approach hinges upon the assumption that California and other states share similar relationships between sociodemographic indicators and pollution outcomes, which may not be accurate in reality, we hope that our mapping tool will aid clean technology companies and policy makers in identifying environmentally at-risk communities and serve as the first step for state environmental authorities and relevant organizations to start collecting (if that applies) and disclosing critical environmental data to the public.

References

[1] P. Brown. Race, class, and environmental health: a review and systematization of the literature. *Environmental Research*, 69:15–30, 1995.

[2] J. Chakraborty, J. Maantay, and J. Brender. Disproportionate proximity to environmental health hazards: methods, models, and measurement. *American Journal of Public Health*, 101:S27–S36, 2011.

[3] L. Clark, D. Millet, and J. Marshall. National patterns in environmental injustice and inequality: Outdoor no_2 air pollution in the united states. *PLoS One*, 9(4):e94431, 2014.

[4] L. Cushing, J. Faust, L. M. August, R. Cendak, W. Wieland, and G. Alexeeff. Racial/ethnic disparities in cumulative environmental health impacts in california: Evidence from a statewide environmental justice screening tool (calenviroscreen 1.1). *American Journal of Public Health*, 105(11):2341–2348, 2015.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[6] P. Mohai, D. Pellow, and J. T. Roberts. Environmental justice. *Annual Review of Environment and Resources*, 34:405–430, 2009.

[7] M. S. ONeill, M. Jerrett, I. Kawachi, J. I. Levy, A. J. Cohen, N. Gouveia, P. Wilkinson, T. Fletcher, L. Cifuentes, and J. Schwartz. Health, wealth, and air pollution: Advancing theory and methods. *Environmental Health Perspectives*, 111(16):1861–1870, 2003.

[8] L. Schweitzer and A. Valenzuela. Environmental injustice and transportation: The claims and the evidence. *Journal of Planning Literature*, 18(4):383–398, 2004.