Mitigating allocative tradeoffs and harms in an environmental justice data tool

Keywords: algorithmic fairness, allocative harms, environmental justice, audit study, racial bias

Extended Abstract

Since 2012, the US state of California has been mandated to redirect 25% of proceeds from the state's cap-and-trade funds to "disadvantaged communities" in order to offset the burdens of pollution to those communities. The California Environmental Protection Agency was thus given the discretion of determining who qualifies as a "disadvantaged community"; to make these determinations, the CalEnviroScreen algorithm was built [1]. CalEnviroScreen takes as input factors like pollution, socioeconomic status, and public health, for each of California's 8,057 census tracts. It then generates a numeric score for each census tract, and the census tracts in the top 25% of CalEnviroScreen scores are marked as "disadvantaged." These disadvantaged communities then receive roughly $525 million annually from California's cap-and-trade funds; CalEnviroScreen is massively financially consequential to the communities who receive the "disadvantaged" designation.

Our work audits the CalEnviroScreen algorithm to determine how and why it can be biased against different communities; we are able to do this by recreating the CalEnviroScreen algorithm based on their released documentation. First, we show that CalEnviroScreen is highly sensitive to small perturbations in the model. We quantify the massive disparities between communities just below and just above the cutoff to receive funding, and showcase the high degree to which CalEnviroScreen can be prone to adversarial optimization (e.g., minor changes to the algorithm can yield favoritism towards specific populations, such as by race or political affiliation). Second, we show that CalEnviroScreen's data processing choices, from normalization methods to feature selection, can disproportionately result in non-disadvantaged scores for truly disadvantaged tracts, especially for tracts with high immigrant populations and tracts with high shares of people of color in poverty. Finally, we discuss the ways in which these harms can be mitigated, through both technical and regulatory solutions.

Recall that CalEnviroScreen is not a supervised learning model; because there is no "ground truth" of what it means to be a truly disadvantaged census tract, we instead focus our evaluations on model ***sensitivity***: i.e., given reasonable and minor changes to the model itself, how much do the output scores change, and to what extent do the set of top 25% disadvantaged census tracts change? Our variations of the model include minor changes of (a) data pre-processing methods, (b) data aggregation, and (c) variable definitions; however, these minor changes lead to very different output scores. We find CalEnviroScreen is highly sensitive: under slightly different variants of the model, the census tracts originally designated as "disadvantaged" (due to being in the top 25% of scores) could vary by 44 percentile ranks. Per Figure 1, overall, over 16% of all census tracts could switch designation (from disadvantaged to not, or vice versa) under slightly different models.

The amount of funding available to census tracts with the "disadvantaged" designation is significantly more than those without. Because we are able to generate the CalEnviroScreen model scores for each census tract, we can identify census tracts just below the 25% cut-off, and just above it. By running a regression discontinuity design, we find that the causal effect of receiving the "disadvantaged" designation is a 104% increase in funding (see Fig 2). This is equivalent to $2 billion over four years for about 2,000 census tracts.

Given the potential monetary incentives to be designated as "disadvantaged," we then explored the extent to which CalEnviroScreen could be adversarially optimized by a decision-maker in control of the algorithm. We model this adversarial optimization to optimize for a specific demographic (e.g., race or political affiliation), and solve using the Hooke-Jeeves method. We find that an adversarial planner could yield a 35% relative increase or decrease for tracts favoring a specific political party by making slight changes to the model (as was done in our sensitivity analysis described above). Similar changes can be made to optimize for or against POC populations with high effectiveness.

We now turn to disentangling how data processing choices can lead to ***fairness*** concerns for specific underserved populations, and in doing so, provide examples of specific slight changes we made to the CalEnviroScreen model. First, we discuss the choice of population health data used as inputs to CalEnviroScreen: one such variable is asthma attack-related ER visits. Prima facie, this seems very reasonable – asthma is a lung disease closely connected to pollution, and ER visits are an apt way to measure severe asthma attacks. However, prior work has shown that immigrants and refugees are both less likely to go to the ER [2], and less likely to have asthma [3]. As such, we propose substituting the input variable of asthma attack ER visits with a different variable: tract-level survey results on chronic obstructive pulmonary disease (COPD) prevalence. When this one variable change is made to the original CalEnviroScreen model, the resulting model variant defines disadvantaged tracts similarly for tracts with low shares of foreign born populations, but yields very different outcomes for census tracts with higher than 30% foreign born populations (Fig 3) – indicating that using asthma as the only indicator of respiratory health is potentially biased against immigrant and refugee communities.

Next, we discuss the choice of a data pre-processing method: standardization. CalEnviroScreen uses percentile ranking to standardize input data. While standardizing data across input variables makes sense – e.g., to ensure that a unit of air pollution is comparable to a unit of traffic – performing percentile ranking instead non-linearly distorts the differences between values. This is especially concerning for variables like PM2.5 mass, since fine particulate matter is generally acknowledged in the public health literature to be linear with respect to health outcomes. A tract with PM2.5 of 11.8μg/m^3 is at the 70th percentile in our data, whereas a tract with a nearly identical PM2.5 mass, 12.3μg/m^3, is at the 90th percentile – a large percentile difference that does not represent the underlying similarity in PM2.5. Meanwhile, a tract at the 100th percentile has a PM2.5 mass of 16.4μg/m^3, which is significantly larger than the raw PM2.5 value at the 90th percentile, but is masked by a relatively small percentile difference. We make one change to the CalEnviroScreen model by performing z-score standardization instead of percentile ranking (Fig 4), and find that over 5% of census tracts change their "disadvantaged" designation, corresponding to $75 million annually changing hands. We can trace this further by looking at changes in race distributions and categorizing the tradeoffs that necessarily occur when alternate models are used. Generally, our proposed alternative model incorporating z-score standardization and COPD survey data resulted in increasing disadvantaged status among census tracts with higher levels of people of color *in poverty*; however, this decreased disadvantaged designation among some tracts with higher populations of people of color *overall* (Fig 5).

Finally, we propose to mitigate these harms by incorporating additional model variants and providing funding to all tracts that are designated disadvantaged by any model – which we find only increases the total number of funded tracts by 10%, but reduces sensitivity by 39%. Ensemble models, tiered funding, or lottery systems could also be considered. We further recommend that an external advisory committee be put in place to foster community-based environmental justice, and enforce accountability of the California Environmental Protection Agency.

References

[1] August, L. et al. CalEnviroScreen 4.0 (OEHHA, 2021); https://oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf

[2] Tarraf, W., Vega, W. & González, H. M. Emergency department services use among immigrant and non-immigrant groups in the United States. J. Immigr. Minor. Health 16, 595–606 (2014).

[3] Iqbal, S., Oraka, E., Chew, G. L. & Flanders, W. D. Association between birthplace and current asthma: the role of environment and acculturation. Am. J. Public Health 104, S175–S182 (2014).

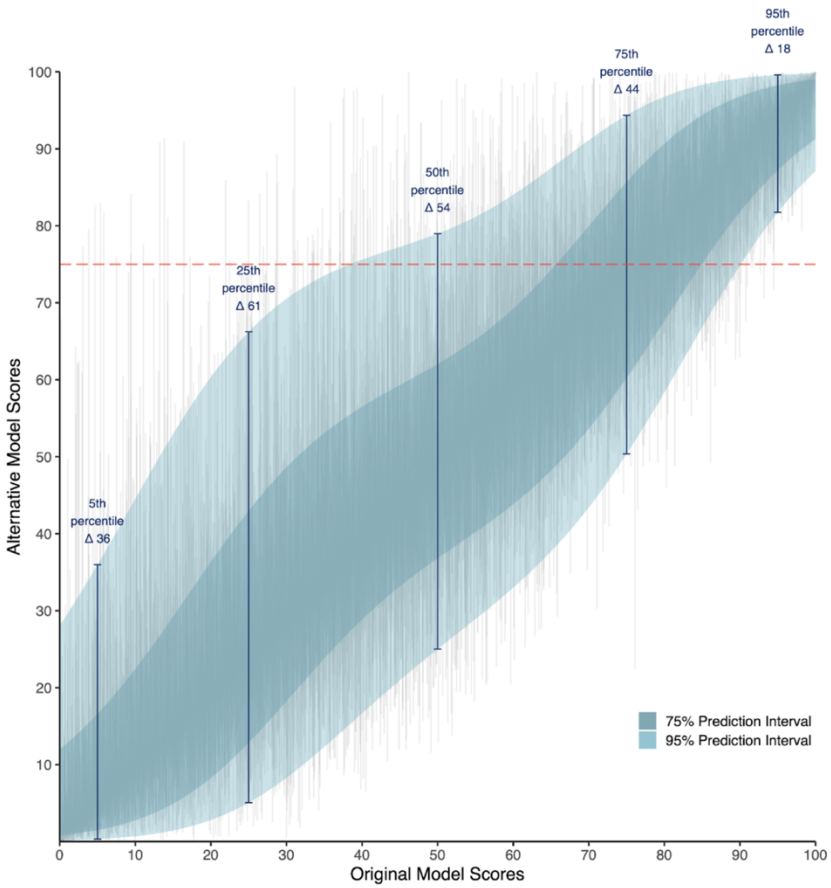

Figure 1. 16% of tracts could change designation under different slightly different versions of the CalEnviroScreen model. Grey bars indicate maximum and minimum values from alternative plausible model specifications with varying health metrics, pre-processing methods and aggregation methods. The horizontal red dotted line indicates the 75th percentile, above which are census tracts that receive the "disadvantaged" designation. Even among the census tracts originally scoring in the bottom 5% of all census tracts (on the left side along the x-axis), some model variants result in scores resulting in the "disadvantaged" designations.

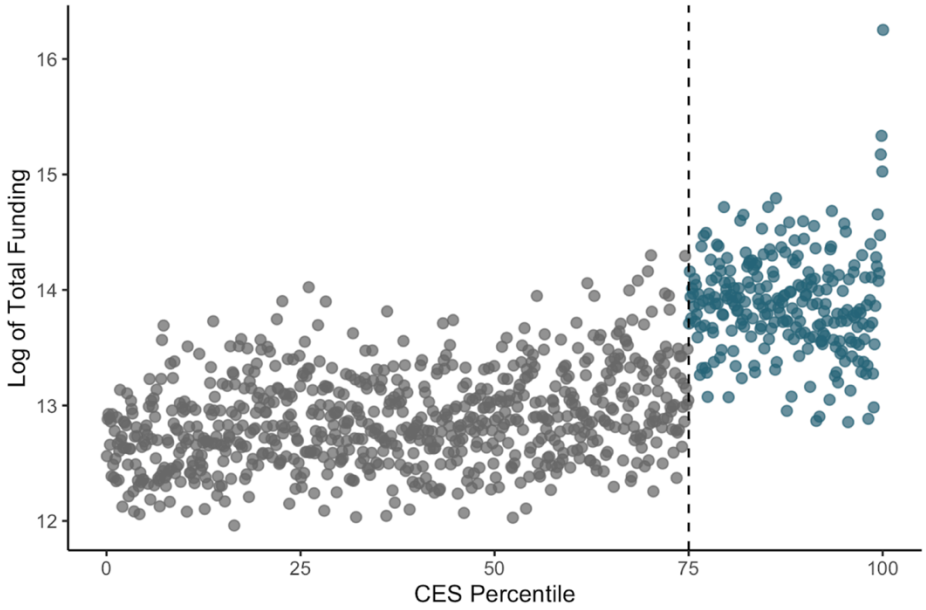

Figure 2. The cut-off for being algorithmically designated as a "disadvantaged" census tract is being in the 75th percentile of CalEnviroScreen scores. A regression discontinuity analysis finds that tracts just above the cut-off are allocated twice as much funding as tracts just below the cut-off.

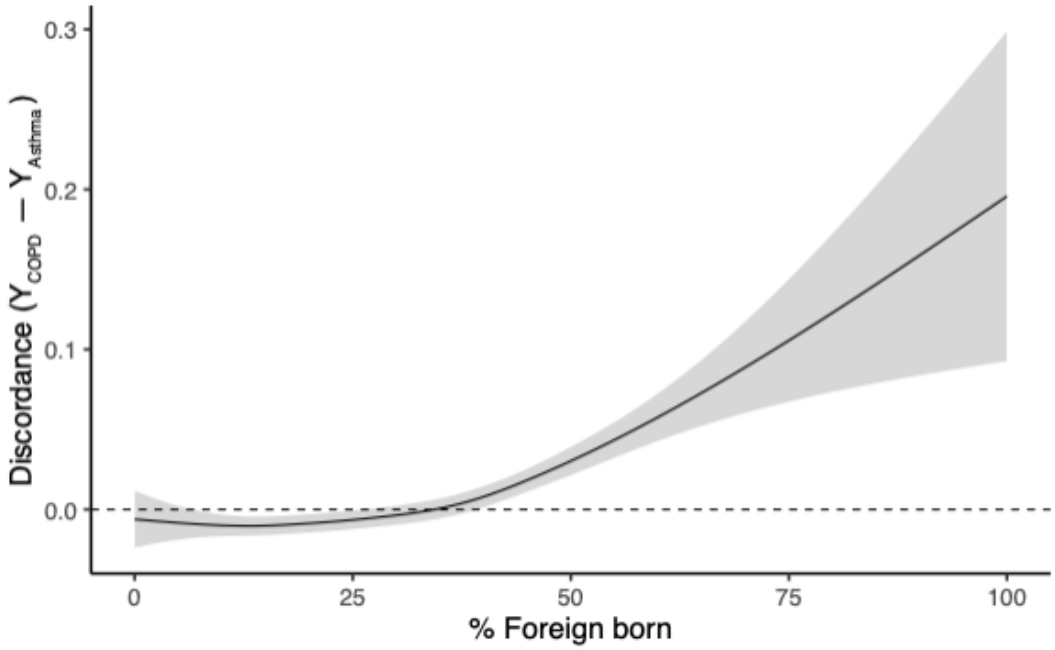

Figure 3. The alternative model (Y_{COPD}) uses survey data of chronic obstructive pulmonary disease (COPD) as a measure of respiratory health compared to the current CalEnviroScreen model (Y_{Asthma}), which uses emergency room visits for asthma. Higher levels indicate Y_{COPD} designating more tracts as disadvantaged for a given foreign born population percentage. Shaded bars indicate 95% confidence intervals, and black line indicates a smoothing spline from pointwise mean estimates of pairwise discordance. The models are comparable for tracts with fewer than a 30% foreign-born population, suggesting model bias against tracts with high immigrant populations.

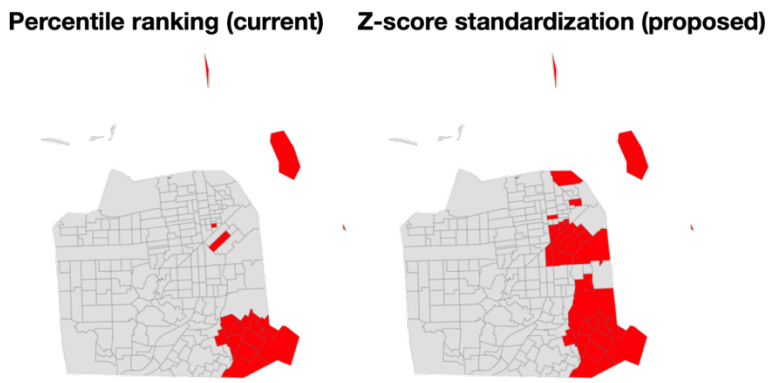

Figure 4. An example in San Francisco: using z-score standardization rather than percentile ranking results in more intuitively disadvantaged neighborhoods (such as the Tenderloin and Chinatown) being algorithmically designated as "disadvantaged".

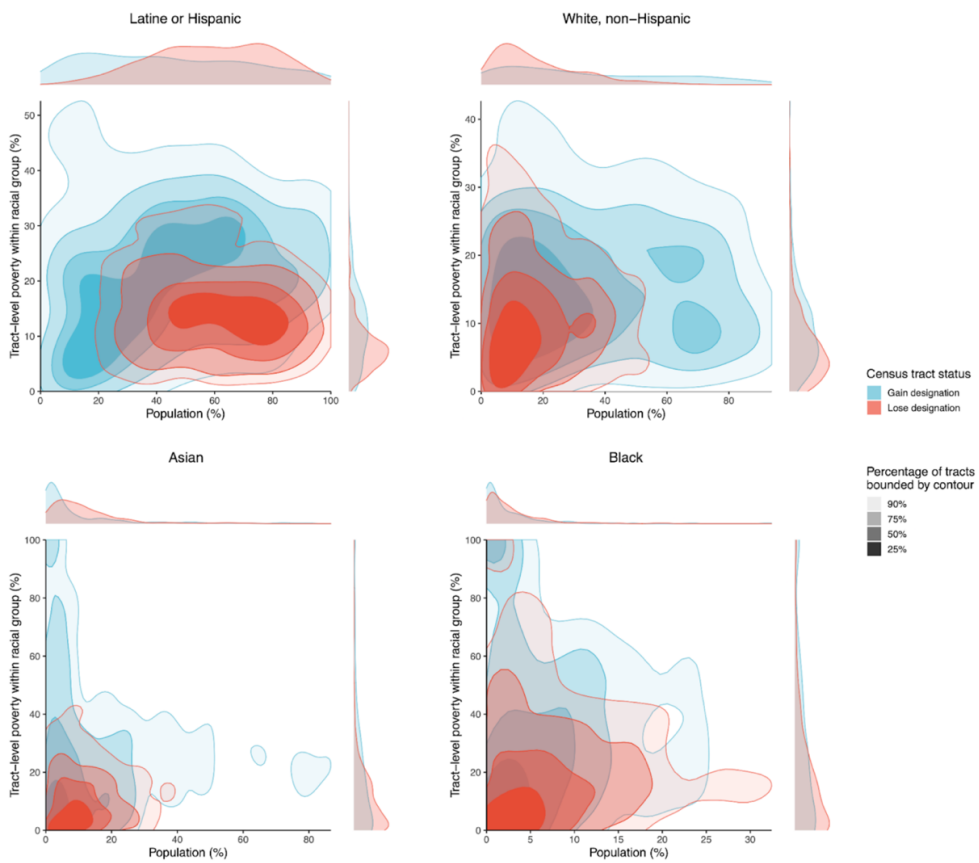

Figure 5. Comparison of how algorithmically designated tracts are distributed by race and poverty across the current and an alternative CalEnviroScreen model. Red densities indicate tracts that receive designation under the current model but are not designated under the alternative model. Blue densities indicate tracts gaining designation under the alternative model. Contours are calculated as the smallest regions that bound a given proportion of the data (highest density region). Dots indicate individual tracts.