# Developing a model to predict income inequalities from air pollution metrics

**Hikaru Hotta**
Department of Computer Science
Stanford University
Stanford, CA 94305
hhotta@cs.stanford.edu

**Cameron Thouati de Tazoult**
Department of Computer Science
Stanford University
Stanford, CA 94305
cameron8@cs.stanford.edu

## Abstract

Environmental discrimination is the disproportionate exposure of low income areas and minority groups to the negative impacts of pollution. Environmental organizations consider it an injustice that those mired in poverty suffer the most from the impacts of pollution despite contributing the least. In our project, we determined the extent to which poverty can be predicted by the levels of air pollution. To do so, we scraped EPA (Environmental Protection Agency) air pollution data by county from 2000 to 2017 and the corresponding household income data from the U.S. Census Bureau. As our baseline, we ran a linear regression on our data-set. Then we implemented two models, a logistic regression and multi-layered linear regression neural network, as predictors of the median income bracket. Both models clearly outperformed the baseline.

## 1 Introduction

Environmental discrimination refers to the disproportionate exposure of low income areas and minority groups to the negative impacts of pollution. It encompasses discrimination in environmental policy making and the enforcement of laws that directly target low income communities. This injustice is frequently observed in the United States. For example, in the early twentieth century, residents of the predominantly Latino neighborhoods of Chicago's Little Village suffered from severe respiratory diseases due to pollutants from nearby coal plants. More notably, since 2014, the residents of Flint, Michigan, a city that is over 55% black and particularly impoverished, have suffered the impacts of lead poisoning from their local water supply[1].

According to the WHO organization, air pollution is directly responsible for 7 million deaths per year[2]. Due to its known impact on living conditions and health, we decided to investigate whether air pollution disproportionately impacts low income communities as a form of environmental discrimination. To do so, we developed models to predict income inequalities from air pollution metrics.

In this paper, we show that it is somewhat possible to predict income distribution from air pollution metrics. We approached this problem by creating multilayered linear regression and multinomial logistic regression neural networks to classify income levels from air pollution data. Our experiments brought mixed results. On one hand, both our models improved upon our baseline. On the other hand, we did not see as large of a spike in the accuracy of our models as expected.

These models are important as they elucidate the relationship between pollutants and income levels. A predictive model for income based on pollution metrics, could potentially be used as a valid metric

[1] https://www.epa.gov/flint/flint-drinking-water-documents
[2] https://www.who.int/airpollution/en/

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

to measure levels of environmental discrimination. This could help inform environmental policy and decisions to negate the injustices of environmental discrimination.

The remainder of this paper is structured as follows: We begin by exploring the state of the art, then delineate our general approach to the problem. Then we outline our key implementation decisions for our models and the experiments we ran on them to optimize their utility. We end the paper with our vision for future work in this field and highlight our most notable findings in the conclusion.

## 2 State of the art

Environmental groups have linked pollution to poverty for some time as it is thought that pollution disproportionately affects low income areas and minority groups (environmental discrimination). A 2015 global review of the socioeconomic disparities in experienced pollution levels [1] found that the majority of literature on environmental inequality from North America, New Zealand, Asia, and Africa has shown that low income communities experience higher levels of air pollutants. However, the study found mixed results from papers published in Europe. The general lack of work on environmental inequality outside of the United States was noted in the review and was attributed to unfamiliarity with the term and its implications. The hope is that as concepts of environmental justice spread globally more work will be done around the world.

Furthermore, very little of the current work on the topic takes a strictly quantitative, computer science approach. Much of the reviewed work simply aggregates other papers or performs basic measurements using publicly available data. Though the current literature acknowledges environmental discrimination regarding air pollution as a measurable phenomenon, as far as we know, ours is the first work to attempt to predict socioeconomic status from air pollution data.

In one related study, John Molitor et al. [2] attempted to identify vulnerable populations in Los Angeles by associating clustered multi-pollutant profiles with socioeconomic status. The authors used a Bayesian framework with Markov chain Monte Carlo methods. Though the resulting relationship was complex and non-linear, the study did generally find that higher levels of pollutants correlated with higher poverty rates. Though this study attempted to quantify pollution related health risks using pollution and poverty data (as opposed to predicting poverty), it provides an interesting framework with which to correlate multi-pollutant profiles with poverty and informed our analysis of our features.

## 3 Approach

### 3.1 Data Collection

We scraped per county air pollution data from the EPA website for the years 2000-2017 which included the features: levels of CO (carbon monoxide), SO2 (sulfur dioxide), NO2 (nitrogen dioxide), ozone, PM2.5 (particulate matter of diameter 2.5 micrometers or less), and PM10 (particulate matter of diameter 10 micrometers or less). We removed certain features from the data that we felt were unhelpful or redundant (e.g. had "CO 2nd max 1-hr" and "CO 2nd max 8-hr" so we removed the 1-hr feature). We were left with 6 features (CO 2nd Max 8-hr, NO2 Mean 1-hr, Ozone 4th Max 8-hr, SO2 2nd Max 24-hr, PM2.5 Weighted Mean 24-hr, PM10 Mean 24-hr) and removed the data for counties that had data available for fewer than 4 of the features. We initially replaced NaN values with the median value of each feature. The air pollution data used FIPS state and county codes which we concatenated to match the county codes present in the median household income data from the U.S. Census Bureau.

#### 3.1.1 Data Exploration

**k-Means** We initially considered running k-means clustering on our features in order to correlate multi-pollutant profiles (clusters) with socioeconomic status as per Molitor et al. [2]. However, after discovering the amount of missing data in our dataset we ran k-means as an informative exercise to find an optimal replacement strategy for missing values. The idea behind this was that if distinguishable clusters were found from running n-dimensional k-means then the appropriate cluster mean could be used to replace missing values. However, our analysis failed to return any significant clustering. A typical 2-dimensional clustering is shown in Figure 1. We performed higher

dimensional k-means as well (as clustering becomes more likely at higher dimensions) but the results were not significantly better.

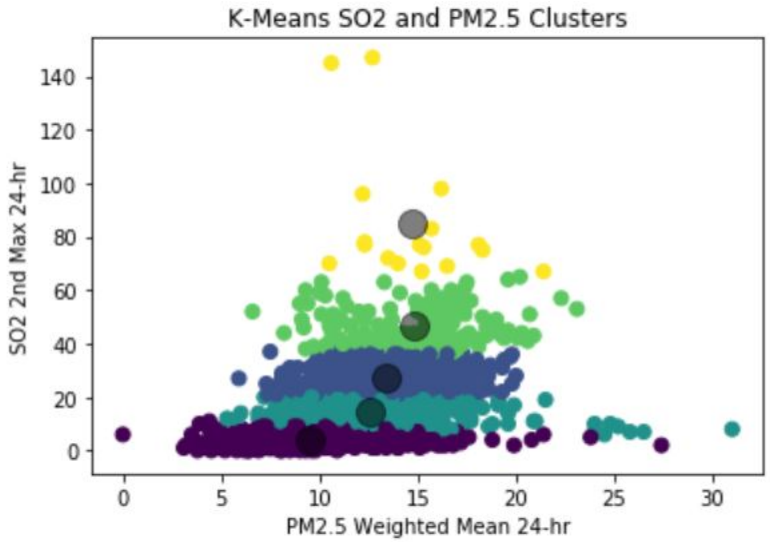


Figure 1: 2-dimensional k-means clusters typical of the results we found. Only 2-dimensions shown for easy visualization.

**PCA** Continuing with our analysis of the current data, we performed a principal component analysis (PCA) of our features. PCA rotates the dataset into a new set of axes so that the new basis is the eigenvectors of old basis. This maximizes variance in every direction and discards useless features (those with no variance), making it easier to classify data. We determined that the feature with the highest variance was PM2.5 followed by CO, SO2, PM10. Beyond simply improving the performance of our model, this analysis was performed to determine which features would be most sensitive to missing-value replacement as well as which components were contributing most to our model. Ozone and NO2 were removed from our feature set following this analysis.

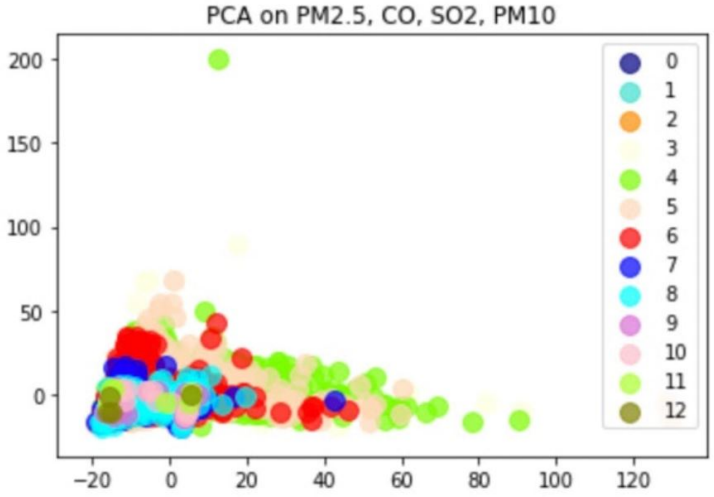


Figure 2: PCA output visualization.

# 4 Key implementation decisions

## 4.1 Baseline

Using sklearn's linear model package, we implemented a linear regression as our baseline. We decided on a linear regression as our baseline because it incorporates a simple linear predictor function.

Our six features were CO 2nd Max 8-hr, NO2 Mean 1-hr, Ozone 4th Max 8-hr, SO2 2nd Max 24-hr, PM2.5 Weighted Mean 24-hr, PM10 Mean 24-hr. Most rows in our EPA data set were missing several values because not all counties measured pollution using all our selected features. Therefore, we replaced missing data with the median index across all counties.

We then split our data into training and test pools and fit our model using X_train, and calculated the mean squared error with X_test, Y_test. Our mean square error was 89580885.98 which was extremely high. However, because we were predicting incomes in the range of several tens of thousands of dollars, this high squared error was expected. Such a value for mean squared error does not mean much qualitatively, so to fix this and obtain more digestible accuracy results we bucketed our income data into 11 income brackets, divided in $10,000 segments. We normalized our data values for each feature by applying feature scaling by using sklearn's preprocessing package to convert our values into z-scores.

We trained our revamped baseline model on the four features, 'PM2.5 Weighted Mean 24-hr', 'CO 2nd Max 8-hr', 'SO2 2nd Max 24-hr', 'PM10 Mean 24-hr' (ozone and NO2 were dropped after running PCA on the data). We allocated 33% of our sample for testing and found that our mean square error was 1.572. This translates to a mean error of slightly more than 1 income bracket. Additionally, we computed our accuracy by running our model on the testing data and computing the percentage of correct predictions from the model. This gave us a predictive accuracy of 0.330.

Table 1: Linear regression baseline coefficients

| Feature | Coefficient |
|---|---|
| PM2.5 | -0.091 |
| CO | -0.143 |
| $SO_2$ | -0.011 |
| PM10 | -0.005 |

## 4.2 Oracle

We implemented an oracle to determine the upper bound we could expect for classification. For our oracle we ran a linear regression similar to that which was used for the baseline, but with "cheating." We added to our set of features population density data (from U.S. Census Bureau). Our final set of features was CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM2.5 Weighted Mean 24-hr, PM10 Mean 24-hr, population density, and housing density (the last two are the new ones). We again replaced missing data with the column median value.

We split the data into training and test pools as we did for the baseline, but our dataset was significantly smaller for the oracle as we only had one year's worth of density data. Our mean squared error this time was 76036269.67 which is still quite high, but significantly lower than the baseline. Re-running the oracle with bucketed data gave us a predictive accuracy of 0.420.

## 4.3 Multilayered Linear Regression

To improve upon our baseline, we decided to develop a multi-layered linear regression neural network. We were curious as to whether adding additional hidden layers to a simple linear regression model would improve our accuracy.

We defined our model as a statistical classification problem with the dependent variable being the income bracket that the median household income belongs to (0-12) and the set of independent variables, our four pollution metrics, which make up the features.

Our model consists of a neural network of three nested linear transformers and rectified linear unit functions followed by a sigmoid function to output probabilities for each income bracket. We again trained our data on the features: PM2.5 Weighted Mean 24-hr, CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM10 Mean 24-hr. In addition, we incorporated a Binary Cross Entropy loss function and Adam optimizer.

The input to the model is a 3328 by 4 matrix of with each row representing a vector of length four containing normalized values for the features: PM2.5 Weighted Mean 24-hr, CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM10 Mean 24-hr. This is a tensor of the form:

tensor([11.2000, 3.0000, 22.0000, 21.0000]).

The target is a 3328 by 12 matrix with each row representing a vector of length 11 containing which indices that denotes the income bracket of the median household income of the corresponding county. For example, a county with a median household income of $50,500 would be in bracket 5, and therefore, the tensor would be of the form:

tensor([0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0])

where only index 5 is on.

After conducting 1000 epochs over our sample, we found our loss decreased from 0.73 to 0.18. The accuracy of our model was 0.404, a significant improvement from our baseline. While this was a noticeable improvement, most of our predictions lay in the range of 3 to 5 which is expected because most counties had median household incomes in those income brackets. Therefore, it is difficult to predict whether our outcome was from our model or from the overall distribution of our data.

### 4.4 Multinomial Logistic Regression

To see if we could improve further upon our multi-layered linear regression, we considered utilizing a softmax function in our model. This is because in theory, discrete data would be better modeled using a logistic regression than a linear regression.

Our multinomial regression model consists of a neural network of three nested linear transformers and rectified linear unit functions. This is followed by a softmax function that computes the probability for each dependent variable.

We again trained our data on the features: PM2.5 Weighted Mean 24-hr, CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM10 Mean 24-hr. In addition, we incorporated a Mean Squared Error loss function and Adam optimizer. Our loss dropped from 0.08 to 0.05 and our accuracy was 0.396. This was an improvement from our baseline, but lower than our accuracy for the multi-layered linear regression model.

Table 2: Accuracy of our three models

| Model | Accuracy |
|---|---|
| Baseline | 0.330 |
| Multi-layered linear regression | 0.404 |
| Multinomial Logistic Regression | 0.396 |

## 5 Experiments

Beyond the data analysis detailed above, we performed several experiments to try and optimize our models.

### 5.1 Replacement-value optimization

Due to the large proportion of missing values in our data (most rows missing at least one of four feature values) we experimented with various replacement strategies in order to best preserve the data and improve it's potential for training our classifier. We initially set missing values to 0, and then

tried setting them to the mean and medians of their respective columns. Additionally, despite the poor clustering, we tried setting missing values to the mean of the nearest cluster returned by our k-means analysis. A comparison of these various strategies showed that setting missing values to the column median gave the best results; slightly better than the mean.

### 5.2 Naïve bayes classifier

To test whether a Bayesian model would be better suited for modelling air pollution as a estimator of income inequality, we implemented a naïve bayes classifier.

Using sklearn's multinomial naïve bayes package, we implemented a multinomial naïve bayes classifier by training our model on 'PM2.5 Weighted Mean 24-hr', 'CO 2nd Max 8-hr', 'SO2 2nd Max 24-hr', 'PM10 Mean 24-hr', and 'Income Bracket'.

In our classifier, we assumed that each feature is independent of each other. Upon testing our model, we achieved an accuracy of 0.356, which was lower than the aforementioned models. This indicates that our naïve assumption may not have been valid. This certainly makes sense since sources of pollution emit multiple types of pollutants. For example, the combustion of fuel emits all of the pollutants in our feature vector. Therefore, it is unlikely that they are independent.

## 6 Future work

Based on the performance of our multilayered logistic and linear regression models, we believe our biggest problem may well be our data. One solution would be to simply use a different dataset such as certain available Kaggle datasets. Another option would be to augment our current data. The majority of our data fell into a few income brackets (with 3 being the most common), making it difficult to train our models. A workaround would be to try and even out the data. For example, we could find whichever income bracket is least common and remove values from the others so that every bracket has the same number of data points. One problem with this is that it would severely decrease the size of our dataset.

### 6.1 Bayesian regression model

One possible development would be the use of a bayesian neural network model. Bayesian neural networks are generally more accurate and, crucially, can perform well with small datasets, a likely outcome if we were to remove certain null or misleading values. This would be particularly beneficial if we end up trimming our training set to even out the number of test samples per bracket. A smaller dataset would also be more vulnerable to overfitting. Bayesian neural networks, however, are fairly resistant to overfitting, making it an intriguing option for further work.

### 6.2 k-fold cross-validation

A limitation of our dataset was the relatively small size (most rows contained at least one missing value). While an obvious strategy for future work would be to acquire and use a larger dataset for training and testing, this may not be possible as the EPA dataset that was used may be the most complete dataset for air pollution in the United States. One solution for better evaluating the models developed in this paper would be to implement k-fold cross-validation. This would be done by shuffling the data and splitting it into k groups, each of which could be used consecutively as a testing set to evaluate a model trained on the remaining data. This could maximize the evaluation potential of our limited dataset to get less biased performance estimates.

## 7 Conclusion

A review of the current literature [1] shows ample scientific evidence for environmental discrimination, particularly in the United States. Though the definition of this term varies, the majority of papers on the subject find that poverty is correlated with higher levels of air pollution. Given that air pollution is one of the most measurable (and historically measured) pollution metrics, it is likely the most available type of pollution data. This makes it a useful component of any quantification of environmental discrimination. We have shown in this paper that a multi-layered linear regression

model using only four prominent features of air pollution can achieve a classification accuracy of 0.404 on a bucketed household income label. This type of model proved to be the most accurate and robust of our analysis. We believe this classification accuracy can be improved significantly in future work through data manipulation and model optimization. Nonetheless, our accuracy already shows marked improvement over the baseline and indicates some level of correlation between air pollution and socioeconomic status. We believe based on our work that the evaluation of machine learning models predicting socioeconomic status based on air pollution can be useful in distilling a quantification of environmental discrimination. Future work with higher accuracy classifiers could give meaningful insight into the complex relationship between air pollutant profiles and poverty and shine light on the root causes of disproportionately high air pollution in low-income areas. More abstractly, we have provided a template for data analysis and modeling when working with multi-pollutant profiles and contributed to the conversation surrounding modeling environmental discrimination.

## References

[1] Hajat, Anjum et al. "Socioeconomic Disparities and Air Pollution Exposure: a Global Review." Current environmental health reports vol. 2,4 (2015): 440-50. doi:10.1007/s40572-015-0069-5

[2] John Molitor, Jason G. Su, Nuoo-Ting Molitor, Virgilio Gómez Rubio, Sylvia Richardson, David Hastie, Rachel Morello-Frosch, and Michael Jerrett Environmental Science & Technology 2011 45 (18), 7754-7760 DOI: 10.1021/es104017x