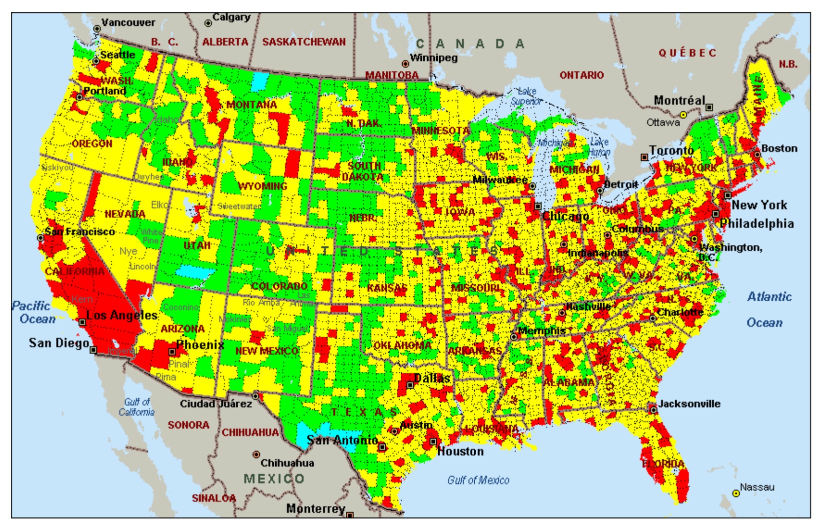
Developing a model to predict income inequalities from air pollution

Hikaru Hotta[1], Cameron Thouati de Tazoult[1]

Background

- **Environmental discrimination** is the **disproportionate** exposure of **low income** areas and **minority groups** to the negative impacts of **pollution**.
- Lower income census tracts have **higher** long-term **$PM_{2.5}$ exposure levels**.
- **Communities of color,** have **higher exposure** rates to **air pollution** than their **white, non-Hispanic counterparts**.
- The **World Health Organization** estimates that **7 million people die** each year from causes directly attributable to air pollution.



Air quality per United States census

Objective

Our objective is to develop a model to predict income level from air pollution levels. This would help us determine the extent to which income inequalities can be predicted from air pollution levels. By doing so, we aimed to elucidate the injustices of environmental discrimination that disproportionately impacts low-income communities.

Data Collection

- Air Pollution Data
  - Environmental Protection Agency – Outdoor Air Quality Data.
  - Features: CO (carbon monoxide), SO2 (sulfur dioxide), PM2.5 (particulate matter of diameter 2.5 micrometers or less), and PM10 (particulate matter of diameter 10 micrometers or less).
  - Data from 2000 to 2017 per United States County.
- Income Data
  - United States Census Bureau - Small Area Income and Poverty Estimates (SAIPE)
  - Household income data per county.
  - Concatenated with Air Pollution Data using FIPS county code

Approach

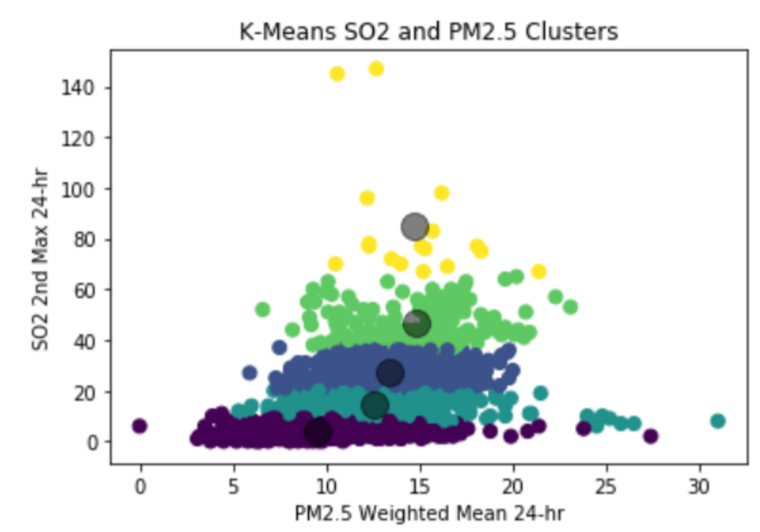
A. Data preparation: Web scraping from EPA and USCB website ➡ Concatenation of dataset by FIPS county code ➡ Filling in of missing datapoints using median values ➡ Feature scaling to convert values into z-scores

B. Data exploration: Feature weighting estimation using scatter plots ➡ Conduct k-means clustering to identify better strategies to fill missing values ➡ Conduct PCA analysis to identify features with the highest variance

C. Model training and validation: Baseline linear regression model ➡ Develop a multilayered linear regression and multinomial logistic regression neural network ➡ Model optimization and validation

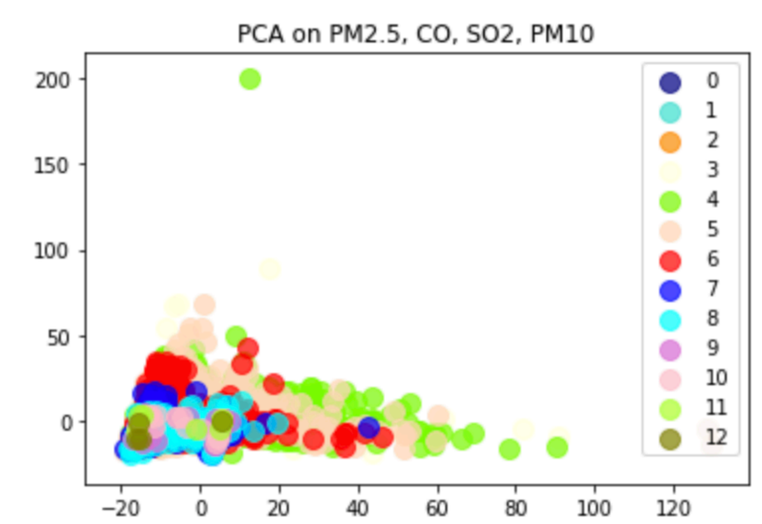
Data Exploration

k-means Clustering



- Ran k-means on our features to identify trends in the data and possible replacement strategies for missing values.
- Unable to get informative clusters.

Principal Component Analysis



- PCA was conducted on all data points.
- Features with the highest variance was PM2.5, followed by CO, SO_2 and PM10.

Models

Multilayered Linear Regression Model

```
nn.Sequential(nn.Linear(n_in, n_h),
                        nn.ReLU(),
                        nn.Linear(n_h, n_h),
                        nn.ReLU(),
                        nn.Linear(n_h, n_out),
                        nn.Sigmoid())
```

- 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.
- Features: PM2.5 Weighted Mean 24-hr, CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM10 Mean 24-hr.
- Binary Cross Entropy loss function and Adam optimizer.

Multinomial Logistic Regression Model

```
nn.Sequential(
    torch.nn.Linear(num_features, num_features, bias=True),
    torch.nn.ReLU(),
    torch.nn.Linear(num_features, num_features, bias=True),
    torch.nn.ReLU(),
    torch.nn.Linear(num_features, num_classes, bias=True),
    torch.nn.ReLU(),
    torch.nn.Softmax(dim=1))
```

- Consists of a neural network of three nested linear transformers and rectified linear unit functions followed by a soft-max function that computes the probability for each dependent variable.
- Features: PM2.5 Weighted Mean 24-hr, CO 2nd Max 8-hr, SO2 2nd Max 24-hr, PM10 Mean 24-hr.
- Mean Squared Error loss function and Adam optimizer.

Results & Error Analysis

Model Accuracy

Model	Prediction Accuracy
Linear Regression Baseline	0.330
Multilayered Linear Regression	0.382
Multinomial Logistic Regression	0.393

- Optimization was conducted by experimenting with different features, additional hidden layers, different loss functions, optimizers, and number of epochs.
- There was a clear improvement from the baseline to both the multilayered linear regression and multinomial logistic regression.
- A higher spike in accuracy was expected for the logistic regression because income brackets/buckets are discrete, which we thought would be better modeled using a logistic regression than a linear regression.
- The abundance of missing data could have contributed to the low learning rate.

Conclusion and Further Work

- We were able to develop two models that improved on the baseline.
- 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 so that every bracket has the same number of data points. This could increase our learning rate and subsequent accuracy.
- 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.
- A random forest classifier could also be an intriguing option as a means to control overfitting and to visualize the decision tree involved in our model's classification.

References

- Bravo, Mercedes, et al. "Racial Isolation and Exposure to Airborne Particulate Matter and Ozone in Understudied US Populations: Environmental Justice Applications of Downscaled Numerical Model Output." Environment International, Pergamon, 23 Apr. 2016, www.sciencedirect.com/science/article/pii/S0160412016301386.
- Molitor, John, and Jason Su. Identifying Vulnerable Populations through an Examination of the Association Between Multipollutant Profiles and Poverty. Environmental Science and Technology, 28 July 2011, pubs.acs.org/doi/pdf/10.1021/es104017x.
- "United States Air Quality." United States Air Quality Map, creativemethods.com/airquality/maps/united_states.htm.