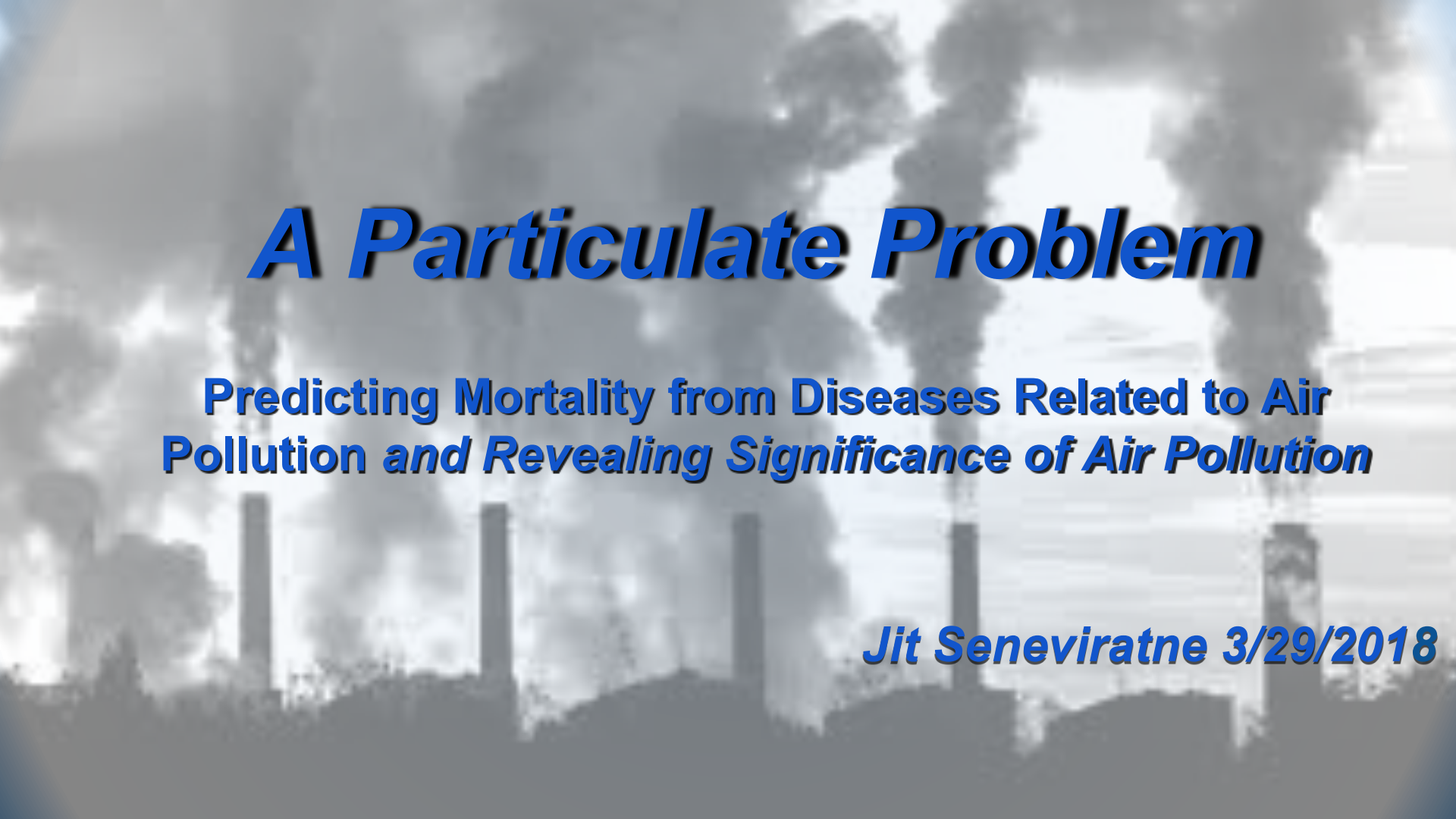


A Particulate Problem

Predicting Mortality from Diseases Related to Air Pollution *and Revealing Significance of Air Pollution*

Jit Seneviratne 3/29/2018

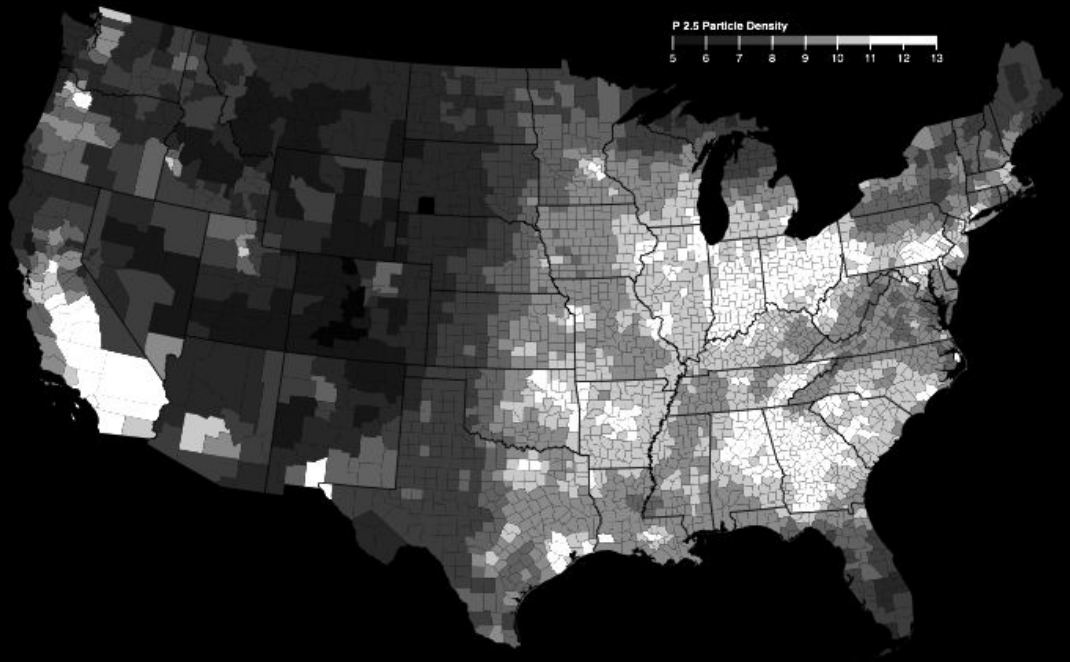
Data Sources

Deaths by IHD and COPD from CDC (per age group)

IHD = Ischaemic Heart Disease

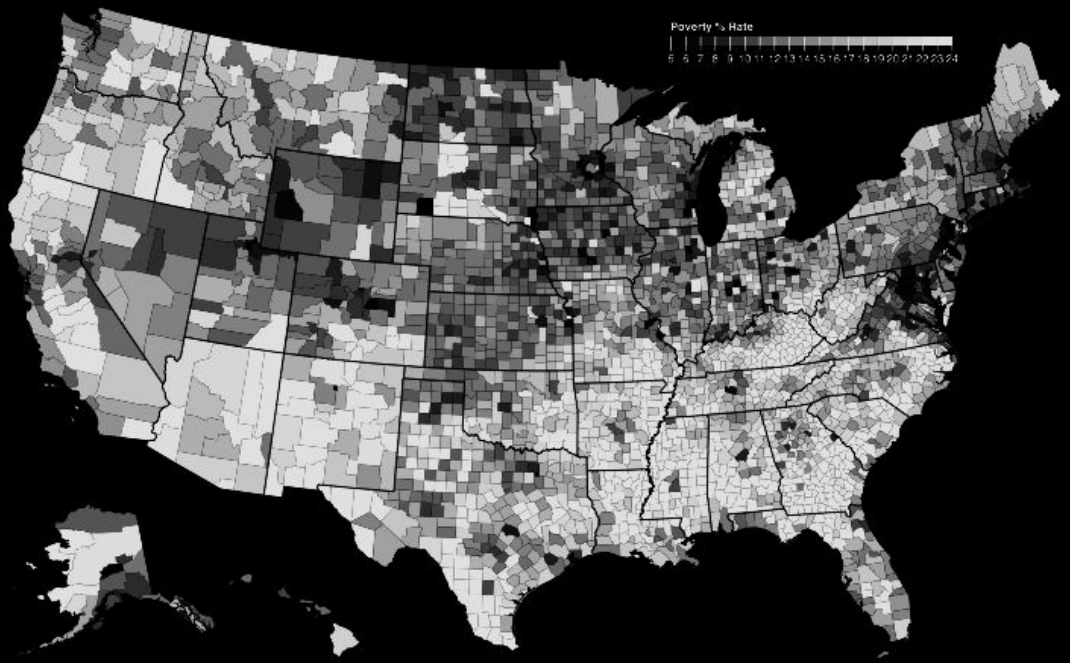
COPD = Chronic Obstructive Pulmonary Disease

Data Sources



Air quality from 2001 to 2011 from EPA using $P_{2.5}$ (fine particle) density

Data Sources



Poverty and Median Income Data from USDA

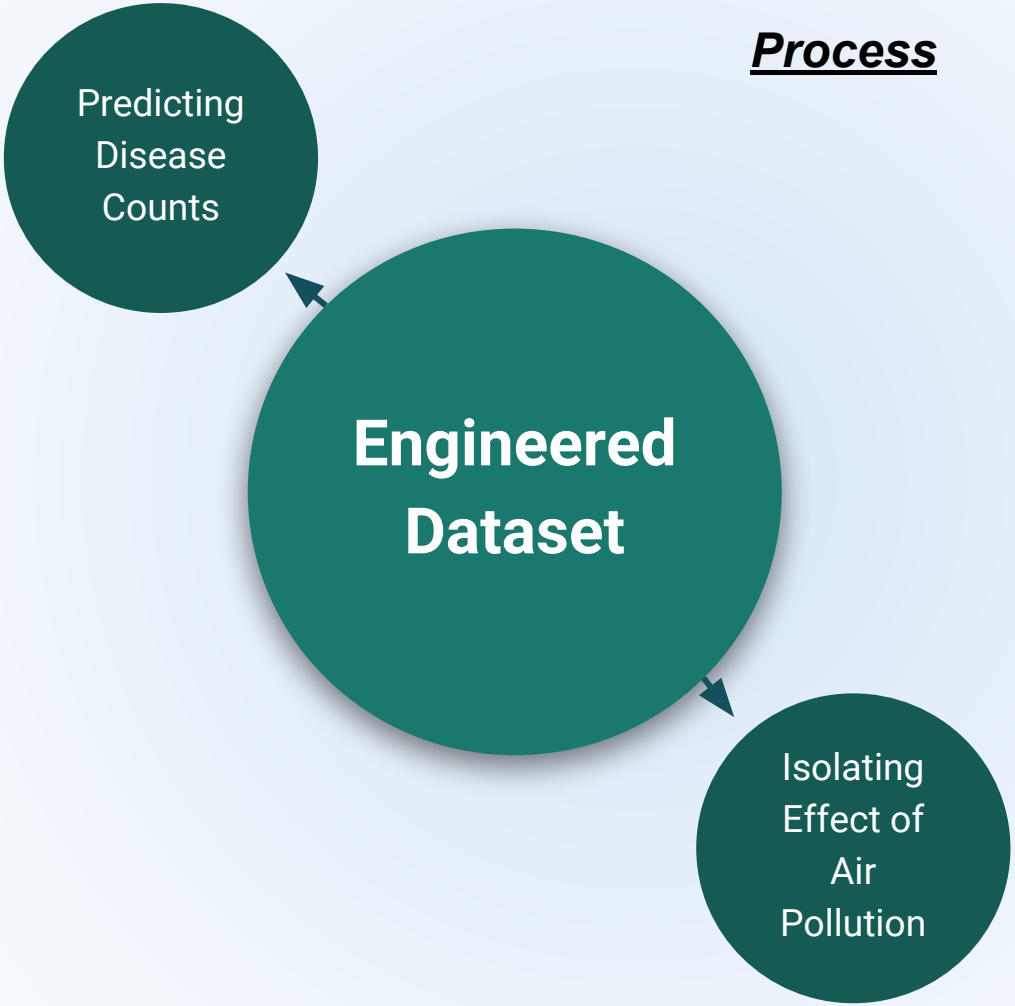
Data Sources

% Diagnosed with Diabetes from IHME

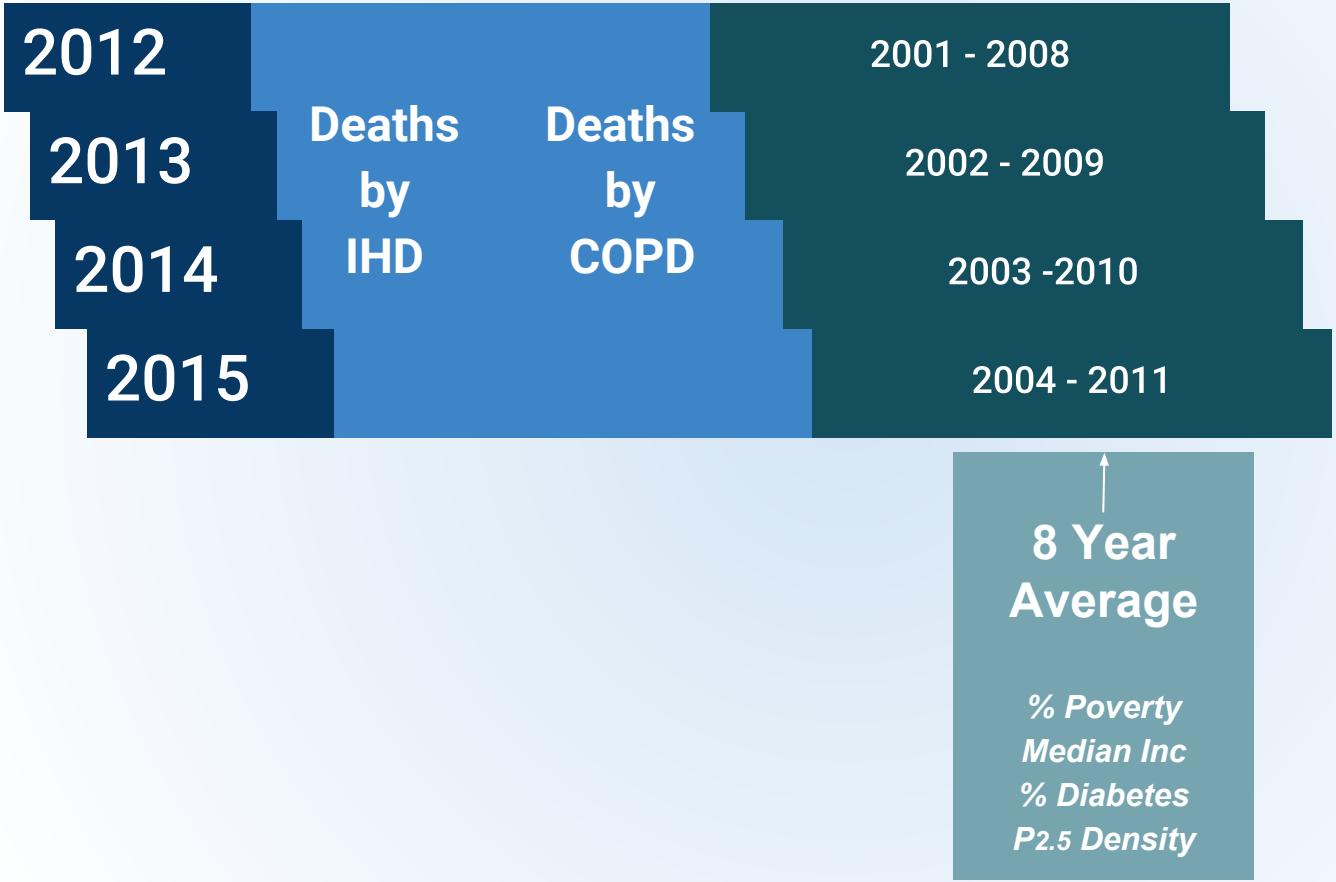
Size of Metro from CDC

Latitude and Longitude

Process


Predicting Disease Counts
Engineered Dataset
Isolating Effect of Air Pollution

Engineered Data



Poisson GLM for COPD

R^2 : 0.69

Rank

1. Population
2. Age Group
3. Diabetes
4. Urbanization
5. Latitude
6. Longitude
7. Poverty
8. Year
9. $P_{2.5}$ Density

Coefficients Unreliable

Poisson GLM for IHD

R^2 : 0.72

Rank

1. Population
2. Age Group
3. Diabetes
4. Poverty
5. Urbanization
6. Year
7. Latitude
8. $P_{2.5}$ Density
9. Longitude

Regular Time Series (One Year Lag on Features)

Poisson GLM for COPD

R^2 : 0.71

Poisson GLM for IHD

R^2 : 0.76

Random Forest Regression

IHD

Test Score	0.89
Train Score	0.95

Rank

1. Age Group
2. Population
3. Diabetes
4. Longitude
5. Latitude
6. Poverty
7. **$P_{2.5}$ Density**
8. Year
9. Urbanization

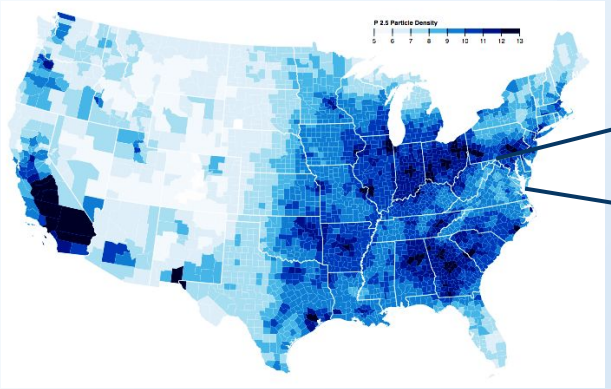
COPD

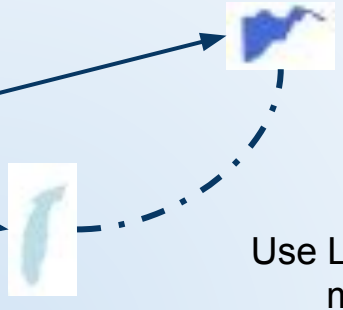
Test Score	0.86
Train Score	0.94

Rank

1. Population
2. Age Group
3. Longitude
4. Latitude
5. **$P_{2.5}$ Density**
6. Poverty
7. Diabetes
8. Urbanization
9. Year

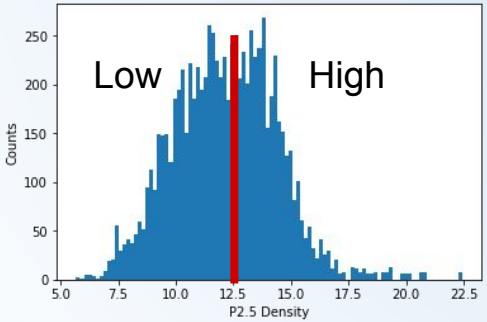
Propensity Score Matching to Isolate Effect of P 2.5 on Deaths





Use Logistic Regression to match pairs on all confounders and yield pairs ***very*** similar in probability

Create ‘high’ and ‘low’ groups based on $P_{2.5}$ density



Propensity Score Matching to Isolate Effect of P 2.5 on Deaths

COPD County	Year	Five Year Age Group	% Diabetes	% in Poverty	Urbanization Code	Deaths / 1000
Camden NY	2012	65 - 69	8.68	10.73	Large Metro	.712
Bristol MA	2012	65 - 69	8.22	10.26	Large Metro	.679

Run for All Pairs

. . . .

. . . .

Run a T-Test for ***mean difference in deaths*** of paired data

Inference on Mean Difference between ‘High’ and ‘Low’ Paired Data using T-Test

Metric	IHD	COPD
Mean Difference in deaths between ‘high’ county and ‘low’ county groups	-0.007	0.22
P - Value	***0.892 X***	***0.02 ✓***
% Diff in Deaths Between Groups	***-0.1%***	***7%***

Outcomes:

- Predictive power of models is strong
- Effect air pollution on COPD deaths is statistically significant
- Effect of air pollution on IHD deaths not significant with current sample size (required sample size for 80% power is 4 million observations)

Thank you!

jitsen.design@gmail.com

www.linkedin.com/in/jitsen

github.com/jitsen-design