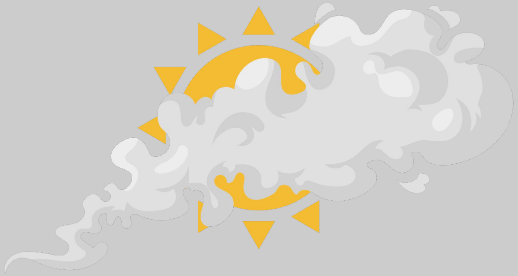
Air Quality and Low Birth Weight

Winston DeGraw, Matt Hurless, Lindsey Hunnicutt

Background



- Air pollution linked with health complications, such as low birth weight (LBW) [1]
    - Risk of birth complications, infant mortality, other health & developmental issues [1, 2]
- **Key pollutants = O_3, Particulate Matter (PM2.5, PM10), CO, NO_2 and SO_2**
- Our group was hired by a task force to shed further light on this issue

Goal: to predict high rates of low birth weight in each U.S. county based on air quality data

1. Maternal exposure to outdoor air pollution associated with low birth weights worldwide. https://deohs.washington.edu/news/maternal-exposure-outdoor-air-pollution-associateMaternal exposure to outdoor air pollution associated with low birth weights worldwided-low-birth-weights-worldwide
2. Low Birth Weight. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5494252/

Key Terms

AQI: (Air Quality Index) index that ranks air quality from good -> hazardous

AQS: (Air Quality System) contains air quality data collected from monitoring sites

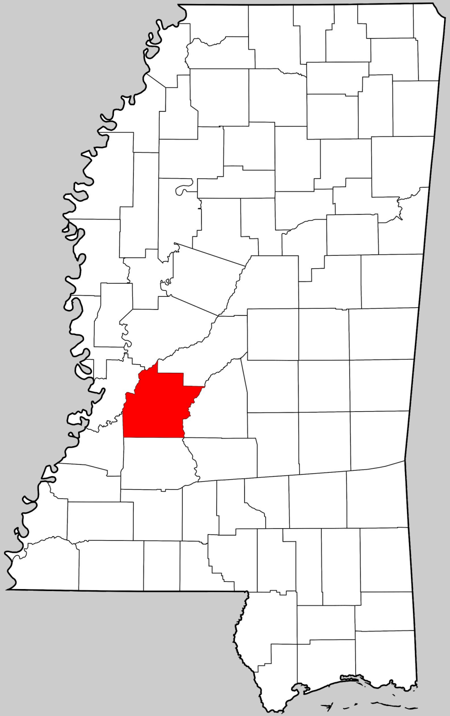
Low Birth Weight (LBW):

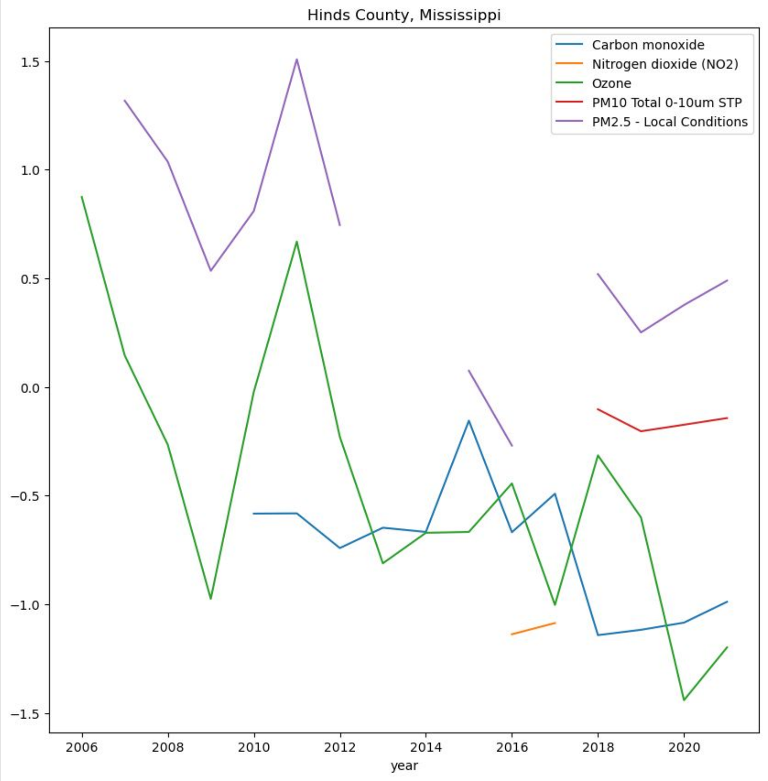
Less than 5 lbs., 8 oz. (2500 g)

Data Collection / Selection

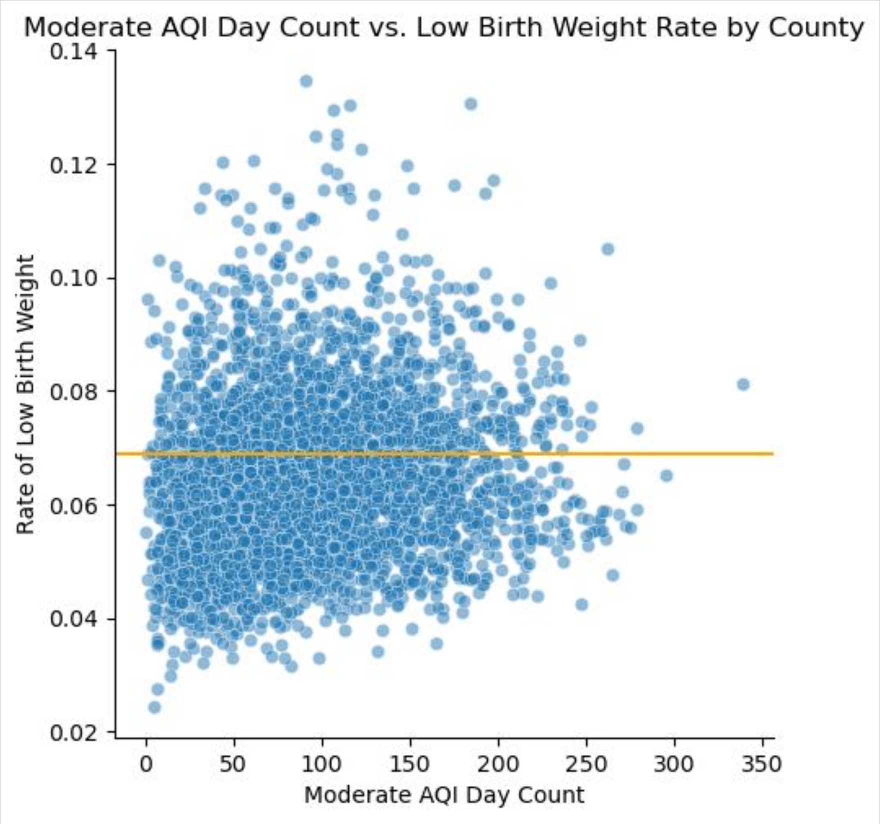
- Data collected from EPA for Air Quality and CDC for birth weight
  - Used the CDC's WONDER system to pull natality information for 2007-2021
  - Used the EPA's Air Quality System (AQS) API to pull airborne pollutant information for 2006-2020
  - Both sets are annual summary data at the county level
- Air Quality data in both AQI day counts as well as more specific sensor data from the AQS API
- CDC data was filtered to exclude any possible maternal risk factors in order to focus on specific impacts of air quality

EDA Insights - Hinds County, MS

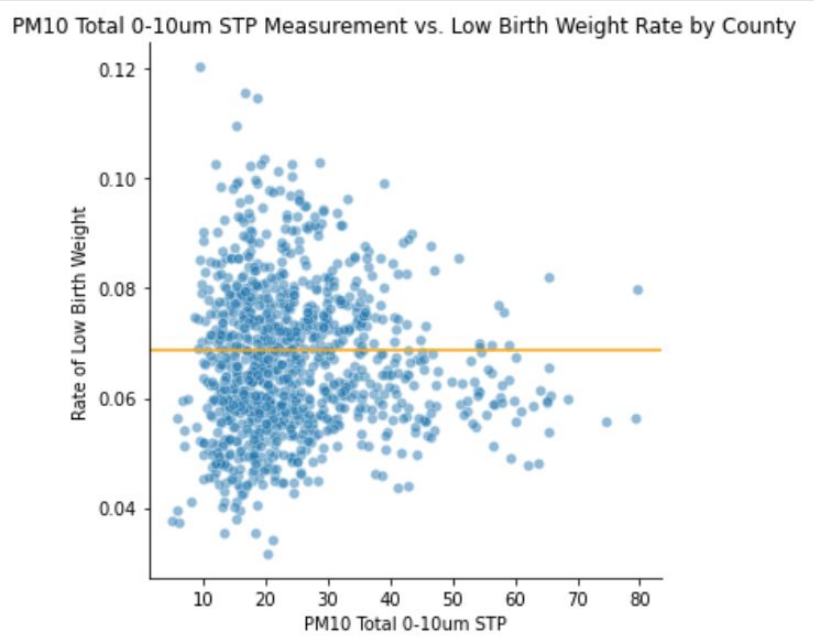


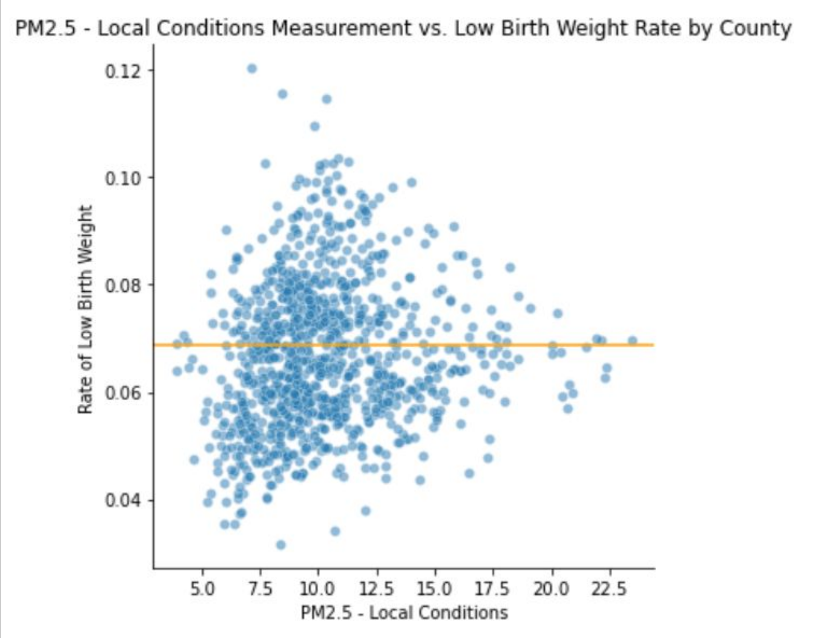


EDA Insights



EDA Insights





Methods

- Build a classification model using air quality to predict excess of low birth rates in the following year
    - High rate = above current national average of 6.88% (calculated after excluding risk factors)
- Multiple models were tried to select a leading model
    - Random Forest
    - Neural Network
    - Logistic Regression
- Required the presence of all measurable parameters for entry to be counted
    - Led to LARGE decrease in the usable data (<7% kept)
    - Wanted to be able to pull feature importance from data if possible

Model & Evaluation

- Neural Net model was only one with predictive power over baseline (58%)
- 17 point improvement in accuracy; 40% of the total improvement possible
- 6 hidden layers with dropout layers
- Each layer used ‘elu’ activation as ‘relu’ did not perform as well
- AdaGrad descent algorithm

	0	1	accuracy
precision	0.789916	0.702381	0.753695
recall	0.789916	0.702381	0.753695
f1-score	0.789916	0.702381	0.753695

Conclusions & Recommendations

- We were able to create a model that can predict LBW rate classification based on air quality data

- Biggest challenge was incomplete data - more robust air quality data is needed

Next Steps

- Attempt to make a more complete data set for pollutants using daily or other data
- Bring a more specific time element to the analysis
- Build a model for interpretability to understand links between specific air pollutants and LBW
- Recommend continued research & efforts to reduce air pollution

Thank you for your time