Agricultural Burn Emissions and Asthma Rates of Nearby Counties

Nick Johnson, Emily Robles, Hailee Schuele, Ci Song

Introduction

Agricultural burning is a cheap and quick way to clear fields before sowing new crops. Although this method results in the emission of black carbon and greenhouse gasses, and has been shown to actually decrease soil health, it is still practiced globally. In the United States, 1-3% of harvested land undergoes burning each year, resulting in greenhouse gasses being emitted from up to 12,000,000 acres of land. Among these gasses, methane is a primary contributor to ground-level ozone, which is detrimental to lung tissue. It's estimated that ozone is the main cause of about one million premature deaths globally each year. We chose to focus our analysis on asthma, a respiratory disease that affects nearly 26 million people in the United States that is believed to be caused largely by environmental factors.

Using data from 2019, we aimed to analyze whether methane emissions from crop fires affect the asthma rates of nearby counties within the United States.

We feel that this information is important for not only state governments to be aware of, but also for the counties that are in proximity of agricultural land that practices crop burning. Additionally, our findings may provide support for the reduction of burning practices, such as in California, which places a limitation on crop burning by designating Permissive Burn Days.

Data

The data for this study comes from two main sources, with additional support from U.S. Census data. The first is crop fire emissions data generated by the Climate Trace project. Climate Trace estimates emissions of greenhouse gases from a variety of industries, including the agricultural sector, and makes this global data publicly available. Climate Trace estimates emissions from cropland fires by leveraging multiple datasets and following a top-down methodology. An asset is defined by Climate Trace as the sum of all emissions for cropland aggregated at the spatial resolution of 0.25 degrees (~25km spatial resolution). Our second dataset is from the PLACES project, which is a joint initiative of the Centers for Disease Control and Prevention, the Robert Wood Johnson Foundation, and the CDC Foundation. The goal of the project is to gather health related estimates at the local level, including by county. For our research we used the asthma rates, smoking rates, and air temperature measurements at the county level. We used county population and land area (sq miles) measurements from U.S. census data to create a population density variable. We chose population density because it intuitively acts as an indicator of urban areas, which represent additional sources of emissions such as transportation.

Our analysis uses methane emissions quantities in metric tons (Climate Trace), county-wide asthma and smoking rates (PLACES), county air temperature (data.world), county population (Census), land area (Census) and population density (population/sq mi) variables. We focused solely on 2019 data as to limit a time series component.

Methodology and EDA

The most granular geographical unit across datasets was county. Asthma, smoking, temperature, and population density were already available at the county-level.

For the emissions variable, each cropland asset was assigned to the nearest county using lat/long coordinates. We divided each emissions value by asset area to standardize exposure across assets of different sizes. For

counties that had more than one asset, emissions were averaged. Ultimately we ended up with average emission/sq mi. After cleaning, there were 2,037 county records with a matching cropland asset.

During the EDA process, we discovered that some variables in our exploratory dataset were highly skewed and decided to log transform population density, population, emissions, and county area. For example, the counties containing Houston and Baltimore had very high population densities. Histograms before and after the log transformation can be seen below.
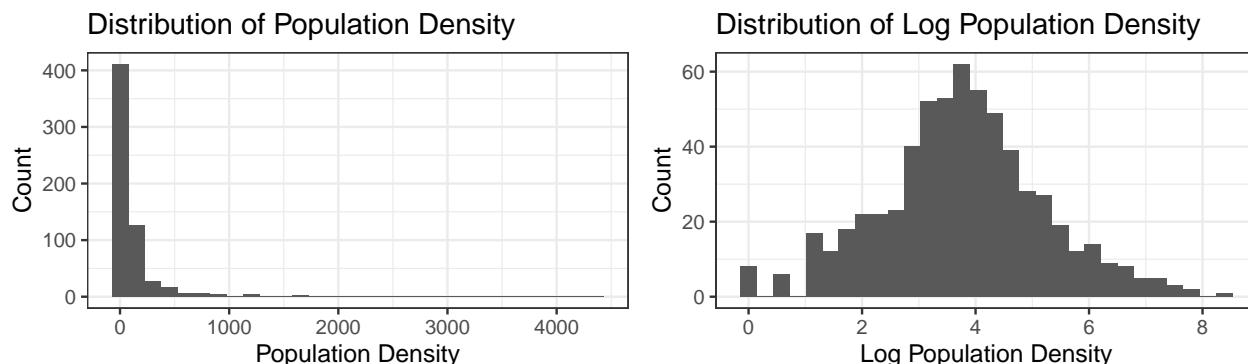
During model creation, we decided that county square mile and population did not make sense as individual predictors. This left us with a few variables, two of which were defined in terms of area; emissions per square mile and population density. We multiplied emissions per square mile with county square mile to create a new emissions variable, which was interacted with population density in the model. This was the closest we could come to a pure emissions value while still being thoughtful about how the two variables were grounded in square mileage and excluding illogical individual predictors. During our EDA, we also noticed a high correlation between asthma and smoking rates (coefficient 0.721).

Variable Correlations in Exploratory Data
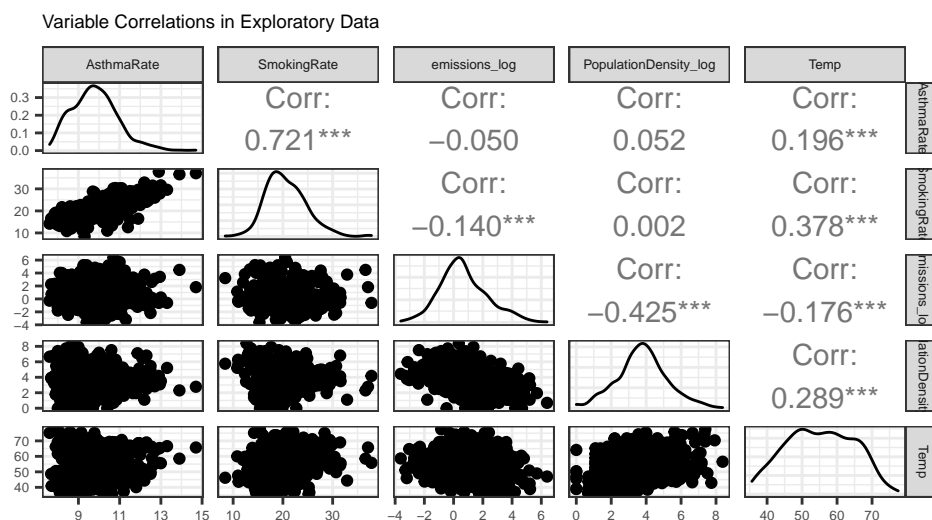
Results

Our first model only included the logged emissions per square mile predictor, which returned an insignificant result with a negative relationship and small effect size. Subsequent models included population and an emissions*square miles interaction, but the considerations discussed in the previous section led us to the final model below.

$$
\begin{aligned}
\widehat{AsthmaRate} = {} & \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot SmokingRate + \beta_3 \cdot Emissions\_log \\
& + \beta_4 \cdot PopulationDensity\_log + \beta_5 \cdot Emissions\_log \cdot PopulationDensity\_log
\end{aligned}
$$

The results below show our initial single coefficient model, an interim model, and our final model on both the exploratory and confirmation data. The results of our exploratory model (30% of data) show significant p-values for all variables. The high R-squared value (0.54) suggests we may have overfitted our model. To take a closer look, we turn to our confirmation dataset. The results of the same model using the confirmation data (70% of data) again show significant p-values and a high R-squared (0.53). What's more, the coefficient estimates are very similar to the exploratory model, albeit with slightly less dramatic effect sizes. The high R-squared values are likely a result of the asthma/smoking correlation mentioned in the EDA section.
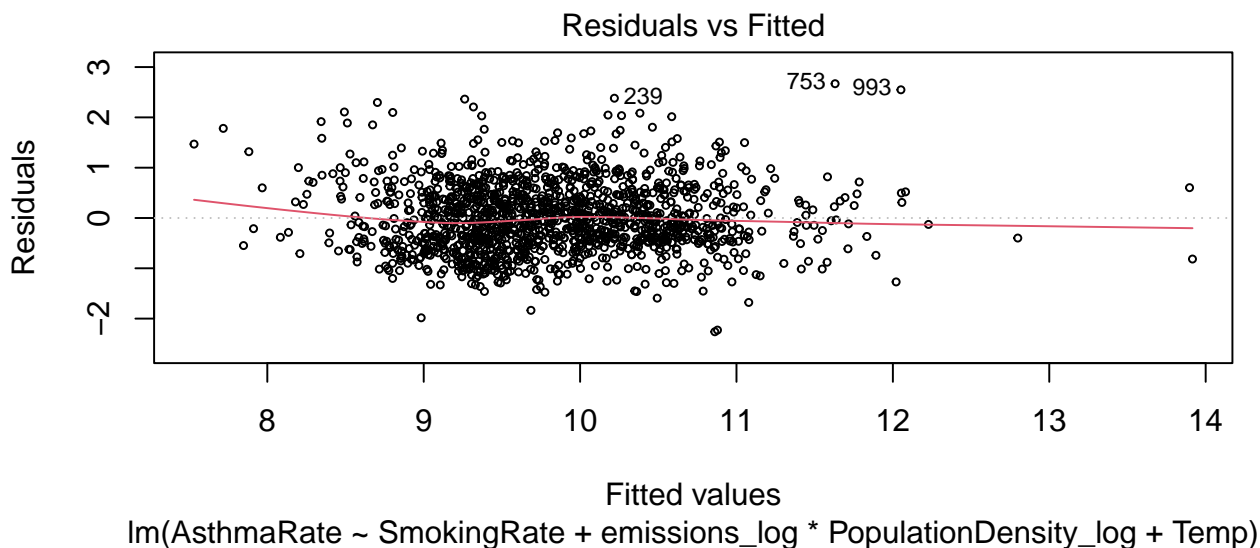
Table 1: Linear Regression Models

	Dependent variable:			
	% Population with Asthma			
	Single Coef Model	Interim Model	Final w/ Exploratory Data	Final w/ Confirmation Data
	(1)	(2)	(3)	(4)
% Smokers in Population		0.215*** (0.009)	0.211*** (0.009)	0.189*** (0.005)
Log of CH4 Emissions Per Square Mile	−0.041 (0.036)	0.017 (0.027)		
Log of Population		0.133*** (0.023)		
Log of CH4 Emissions			0.172*** (0.051)	0.133*** (0.032)
Log of Population Density			0.116*** (0.023)	0.108*** (0.014)
Average Temperature		−0.015*** (0.003)	−0.012*** (0.004)	−0.012*** (0.002)
Emissions & Population Density Interaction			−0.032** (0.013)	−0.018** (0.008)
Constant	9.560*** (0.221)	4.944*** (0.303)	5.600*** (0.274)	6.054*** (0.164)
Observations	611	611	611	1,426
R^2	0.002	0.549	0.544	0.531
Adjusted R^2	0.001	0.546	0.540	0.530
Residual Std. Error	1.099 (df = 609)	0.741 (df = 606)	0.746 (df = 605)	0.682 (df = 1420)
F Statistic	1.471 (df = 1; 609)	184.371*** (df = 4; 606)	144.111*** (df = 5; 605)	321.967*** (df = 5; 1420)

Note: *p<0.1; **p<0.05; ***p<0.01

The plot below shows that the residuals of the confirmation model are evenly distributed, meaning we can obtain unbiased estimates of the regression coefficients.

Residuals vs Fitted

Fitted values
lm(AsthmaRate ~ SmokingRate + emissions_log * PopulationDensity_log + Temp)

Limitations

Limitations exist at multiple levels of our analysis. Satellite sensors used at the highest level of ClimateTrace's estimation methodology detect fires at a minimum size of 375m x 375m and capture data every 24 hours. Small fires and fires lasting a short period of time are therefore not captured by the satellites. Additionally, smoke or clouds can obscure fires from satellites. These limitations suggest a possible underestimation of emissions.PLACES data variables were filtered to only include counties within 15 miles of an asset centroid, meaning that we may have excluded additional exposed counties if they fell only slightly out of our limits.

Evaluating data at the county level also introduces the possibility of geographical clustering, meaning that the data would not meet the assumption of independent and identically distributed (IID) observations required for consistent regression estimates. Reverse causality should also be considered as a limitation of our study.

If pollution is especially bad in a certain area, farmers may be pressured or even incentivized by a county or state to reduce emissions. In this scenario, it is possible that asthma rates would have an effect on emissions by leading to their reduction.

It's worth quickly exploring the interaction between emissions and population density. The plot below shows emissions predictions for the first and third quantiles of densities from the confirmation data. For both populations, emissions lead to higher rates of asthma. However, we can see that interaction term moderates the emissions effect by reducing the slope of the regression line as density increases. We understand this to mean that emissions in urban areas have a similar effect to crop burning and thus dilute the effect of crop fire emissions.

Asthma Rate vs. Emissions for Different Pop Densities
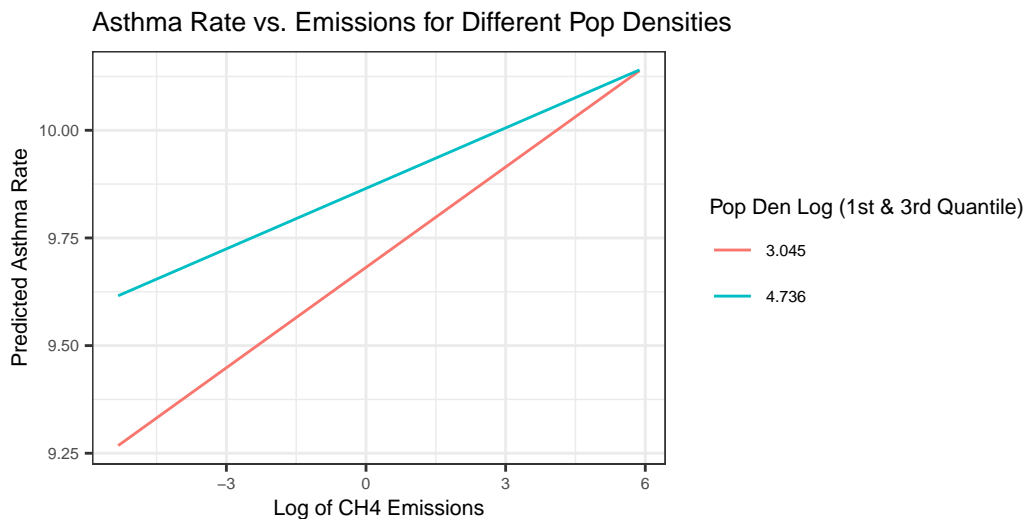
Lastly, while irritation from ozone can be considered an environmental factor that leads to asthma, there are many other contributors that were not included in this study such as genetics, allergies, chemical irritants, and respiratory infections. By acknowledging these limitations, we can better interpret our findings and identify areas for future research to improve our understanding of this issue.

Conclusion

According to the confirmation data model, all predictor variables were significant. The largest estimated effect is the ratio of smokers in a county. For each percentage point increase, there's a 0.19 percentage point increase in asthma rates, meaning roughly 1/5 of those people will get asthma. Temperature presented a modest effect, with warm climates seeing slightly higher rates of asthma. For our primary variable of interest, emissions, we saw a 0.13 percentage point increase in asthma rates for every logged additional metric ton of CH4 emissions. Given that the median log of emissions in both the exploratory and confirmation data is ~0.5 and the maximum ~6, we conclude this is not an extreme quantity of emissions and therefore the effect size is practically significant.

The significance of our findings provides support for local and state governments looking to increase regulation of crop burning. Future research could incorporate additional respiratory diseases beyond asthma, include meteorological factors that affect the dispersion of emissions, or consider socioeconomic features of the nearby counties.