Modelling the Association Between Air Pollution Sources & Health Outcomes on the County Level

Background

Air pollution is a major issue globally, with a recent WHO report (World Health Organization, 2018) attributing 7 million deaths annually to poor air quality, as well as 24% of adult deaths due to heart disease, 25% of adult stroke deaths, 43% of adult deaths due to chronic obstructive pulmonary disease (COPD), and 29% of adult lung cancer deaths. While the US has better air quality than much of the rest of the world, A study from 2013 (Caiazzo et. al, 2013) found that air pollution accounted for approximately 200,000 premature deaths per year (2005 data). In 2005, a total of 2,448,017 deaths were registered in the US, which makes air pollution responsible for about 8% of the deaths in the US that year. In fact, road-transportation air pollution was estimated to account for 58,000 deaths annually, while in 2005 approximately 43,500 people died in auto accidents, making auto exhaust a more likely cause of deaths than an auto accident.

Air pollution can cause mortality in a number of ways, and can affect a number of organ systems aside from the respiratory tract. Air pollution can cause chronic respiratory diseases such as COPD, as well as asthma. It can also negatively affect the cardiovascular system, increasing risk of heart disease and stroke. Air pollution also contains a number of carcinogens which can increase the risks of cancer, especially in the respiratory tract such as in the trachea, bronchus, lung, and esophagus.

The goal of this analysis was to model the association between exposure to sources of pollution. And health outcomes. For health outcomes, two metrics will be used. The first is all-cause age-adjusted mortality, and the second will be a composite unit composed of a number of air-pollution associated outcomes by county, converted. To z scores and averaged.

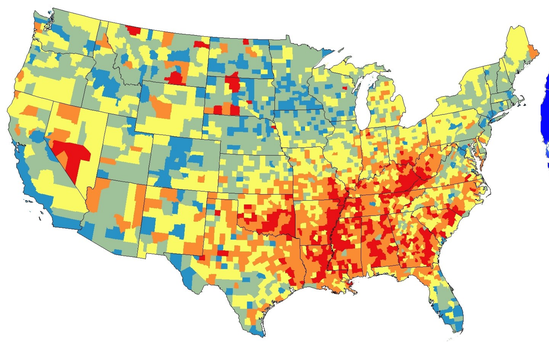
For air pollution exposure metrics, the natural log of county-level highway density will be used to estimate auto emissions exposure, while three raster layers will be used to model exposure to emissions from fossil-fuel power generation, one for coal plants, one for natural gas plants, and one for oil plants. The model will be adjusted by US region (Northeast, South, Midwest, or West), median income, and whether the overall population density of the county is 500 people per square mile, the cutoff for an urban area.

Methods

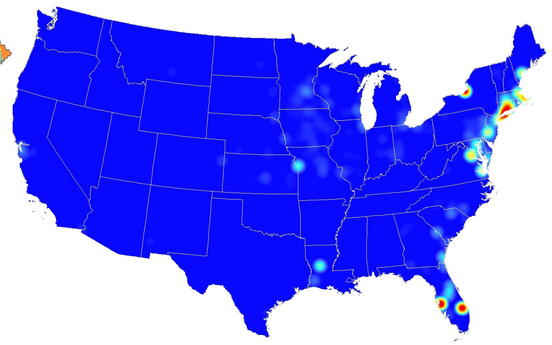
A major and minor highway linefile was projected with a county shapefile, then clipped to determine the miles of highways in a county. County area was used to calculate miles of highway per county, which was then natural-log transformed. Next, a power plants pointfile from the US Energy Information Administration was used to generate three raster layers, one for coal, natural gas, and oil powered power plants. These used a 50 mile radius and the population was defined as the amount of power being generated by the energy source in question. The mean value of each raster was then calculated for each census tract and joined with the shapefile. US regions from the US Census Bureau, Age-adjusted mortality from the Global Health Data Exchange, and median income from US FactFinder were imported for each county.

In excel, 8 air-related health outcomes (chronic respiratory diseases, COPD, asthma , myocardial infarction, cardiovascular disease, ischemic heart disease, ischemic stroke, tracheal bronchial and lung cancer, and esophageal cancer) were combined into one table, converted to Z scores by county, then the mean Z for each county was calculated. These were then merged with the shapefile in ArcMap and exported to Stata.

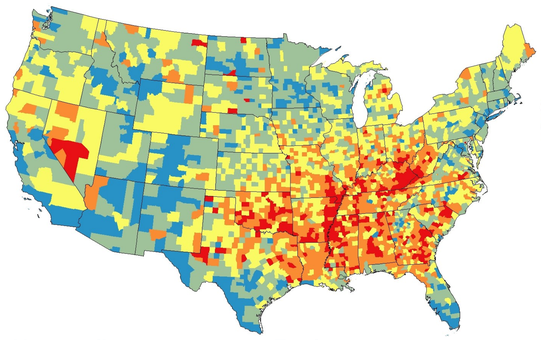
In Stata, four regressions were carried out, two unadjusted and two adjusted. The unadjusted regressions compared age adjusted mortality and the mean Z value (separately) to the three power plant rasters and ln(highway density). The models were then adjusted by US geographic region, household median income, percent white, and by whether the population density of the county exceeded 500 people per square mile (population density was attached to the original county shapefile). The z-value model was additionally adjusted by proportion of the population in each age cohort. This adjustment was not applied to the age adjusted mortality model as age adjusted mortality is already adjusted by age.



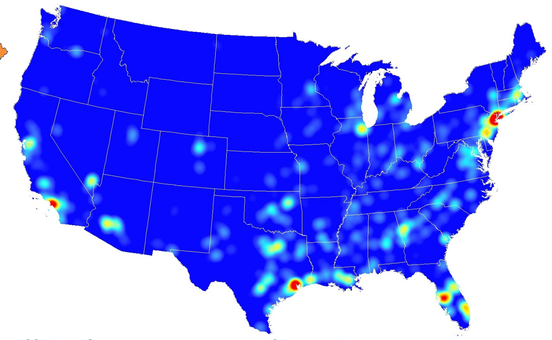
Age adjusted mortality



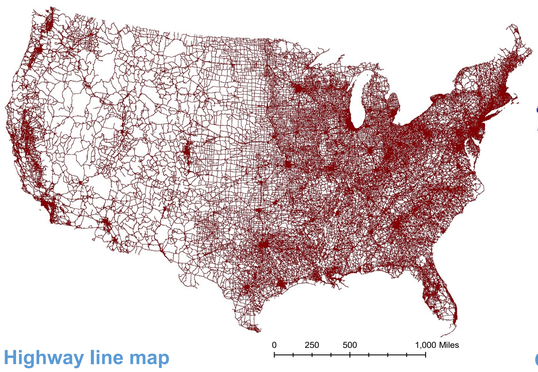
Crude oil power generation raster



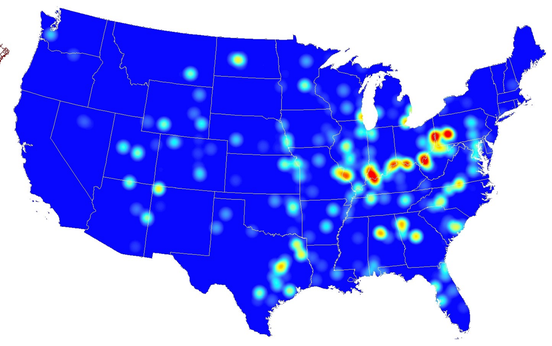
Air mortality outcome mean Z values



Natural gas power generation raster



Highway line map



Coal power generation raster

Model comparisons

	Age adjusted mortality, unadjusted	Age adjusted mortality, adjusted*	Z– value, unadjusted	Z-value, adjusted**
Adjusted R squared	0.0798	0.6604	0.0917	0.6370
Root Mean Square Error	128.72	77.911	0.68465	0.43006
Insignificant variables	Natural gas raster mean	Natural gas raster mean, geographic region = Midwest	————————	Geographic region =West, Natural gas raster
Mean and max VIF (variable)	1.32, 1.47 (Natural gas raster)	2.47, 5.41 (geo. region = South)	1.32, 1.47 (Natural gas raster)	2.84, 5.57 (geo. Rgion = South)

n = 3379 for unadjusted, 3232 for adjusted

* = Adjusted by geographic region, household median income, whether the population density is over 500 (binary), percent smokers, and population percentage white

** = Adjusted by all the same measures as (*), PLUS adjustment by proportion in each age category (under 18, 18-29, 30-39, 40-49, 50-64, 65 and up)

Analysis

Both models were similarly successful in their ability to explain the variability in the data, as shown by similar R squared values of 0.66 for the ageadjusted mortality and 0.64 for the Z metric. While the Z-value measure has a much smaller root mean square error, the scales differ between it and age adjusted mortality, making them incomparable by this metric.

Initially, a Z (air pollution outcomes) minus Z (county death rate) was analyzed, but that was unsuccessful. The majority of air pollution mortality routes are also significant causes of deaths of old age, so the resultant layer displayed places where people tended to die of outcomes that primarily affect the young.

The raster layers revealed some very interesting findings. Of the three fossil fuels, coal was the only one consistently linked with worsening health outcomes. This aligns with existing research showing that coal releases more pollutants than other fossil fuels. The crude raster consistently reported a negative correlation, which was unexpected. I suspect that this is due to the fact that crude raster hotspots tend to be located near large cities (likely for surge power generation), which could account for some of the trend as high population density seemed to have a preventative effect in the model. The other air pollution exposure variable, highway density, was found to be significant in all models, confirming its usefulness as a measure.

Shortcomings and Limitations

1. Counties are a really large unit for exposure, especially when it comes to highway pollution exposure. Better resolution data (say on the census tract level or of individuals) are needed to overcome this.
2. The urban metric is very crude. Almost every census tract has at least one urban center, so the vast majority of urban centers were not included, resulting in a significant proportion of the urban population not getting classified as urban. Improved data resolution could be a fix.
3. The use of aggregate data makes these findings difficult to apply to any individuals risk, and is unsuitable for handling variables such as sex.
4. It is important to remember that the exposure measure in question in not exposure to air pollution; rather it is exposure to air pollution sources. This is a proxy measure and therefore is an inherently imperfect measure of the counties actual exposure to air pollution

Future analysis

Ideally, this analysis would be re-done using a more granular geographic unit, such as census tract. It would also be prudent to find actual air pollution estimates to compare these results against, however that would likely need to be done on a regional basis due to a lack of air quality data, particularly in rural areas. Lastly, a better measure of air pollution deaths could still be made. Doing so would require the population strata and the mortality rate for each of the outcomes and each of the age groups, however with these one could calculate a predicted number of deaths for these outcomes, which could then be used to calculate the “air pollution outcome related deaths” excess or deficit. This measure would overcome one of the major issues with the Z score measure: its lack of age adjustment.

Citations

1) World Health Organization. (2018, May 2). 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. Retrieved May 8, 2018, from http://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

2) Caiazzo, F., Ashok, A., Waitz, I. A., Yim, S. H., & Barrett, S. R. (2013). Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005. *Atmospheric Environment, 79*, 198-208. doi:10.1016/j.atmosenv.2013.05.081

Data sources

County shapefile (includes population, age, race), highway linefile: Tufts GIS data server

Age-adjusted mortality, mortality by health outcome data: Global Health Data Exchange (http://ghdx.healthdata.org/us-data)

Pollution point source data: US Energy Information Administration (https://www.eia.gov/maps/layer_info-m.php)

Median income data: American factfinder

Made by Robert Sucsy

Tufts UNIVERSITY