Jeremy Neiman

**Data Sources and Data Cleaning**

Four datasets were used – Vulcan CO2 emissions data and three datasets from the U.S. Census with county population, county area and a mapping from counties to their MSA (Metropolitan and Micropolitan Statistical Areas) if any. The Vulcan dataset contains county level carbon emissions from 2002 for a number of different sectors, as well as total emissions for the county. The data was originally provided as an Excel spreadsheet. Excel was used to remove comments and convert it to CSV for analysis. The Census came from three different Excel files with a number of fields and comments. The relevant data is the 2002 county population, area and MSA to county mapping. Excel was used to remove extraneous fields and comments and to convert to CSV. The carbon, population and land area datasets contained the county FIPS (Federal Information Processing Standard) code as a five digit code, while the MSA dataset contained them as a two digit state and three digit county code, so these two columns were merged into one. Then the four datasets were merged on the FIPS code. Several counties had zero area in the dataset, so these counties were ignored. For analysis at the MSA level, counties were aggregated by the MSA code. For both individual counties and MSA aggregates, several fields were derived from the given data including CO2 emissions per capita, CO2 emissions per square mile, and population density per square mile.

**Analysis**

There are two primary questions we have. First, how does the total population of a city affect its emissions? Are sprawling megacities more or less efficient than smaller cities? Second, how does the population density affect the emissions? Will a city of a million people emit more or less than another city of a million which is half the area?

First the analysis was done at the most granular level available, the county level and the results are shown in Figure 1. The total CO2 emissions were graphed versus total population of the county and CO2 emissions per capita versus population density of the county. Then ordinary least squares (OLS) linear regression was done to measure any potential correlation. The results are not surprising – they show that total CO2 emissions grow as population grows and that CO2 emissions per capita decreases as population density increases. This analysis was repeated at the MSA level and the results as shown in Figure 2 are the same.

More interestingly is that both total emissions versus population and emissions per capita versus population density are sublinear relationships. This means that as the population grows total emissions will flatten out and as population density grows emissions per capita will flatten out. Figure 3 shows this effect. As shown, the returns on increasing density very quickly disappear and past 20 to 30 thousand people per square mile there is very little gain for increasing density. For example, this model predicts that to cut emissions per capita in Manhattan in half, the population per square mile would need to be increased from about 60,000 people today to about 4,000,000 people per square mile, which would equal over 100,000,000 people in Manhattan. On the other hand, because the emissions appear to go up sublinearly as the city’s total population increases, a city could still grow wide even if there are only negligible gains to growing denser.

To test this hypothesis that there may be an optimal city density but not city size, we can look at the correlation between city land size and CO2 emissions per capita. Since we’re interested in cities that are fairly dense to see if there is any effect on them growing wide, we’ll constrain the regression to counties and MSAs with population density among the top 25th percentile of population density. Figure 4 shows the results. At a 5% confidence level there is no correlation between the area and the CO2 emissions per capita. This would indicate that a dense city could continue to expand in area and a city’s population could continue to increase even once it no longer is fruitful to build the city more densely.

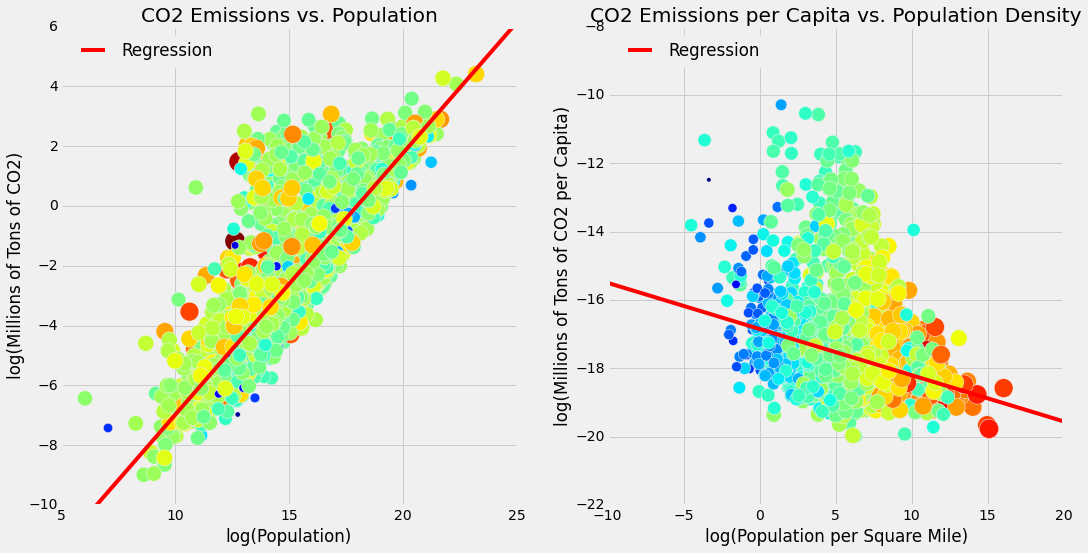


Figure 1: County level log(Total Emissions) vs. log(population) (left) and vs. log(population density) (right). In the left graph point size and color is based on the county’s land area with larger/redder points indicating larger area and in the right size and color is based on total population of the county with larger and redder points having a higher population. Red line shows the OLS regression. For the left graph the slope of the line is 0.8755 and for the right it’s -0.1351.

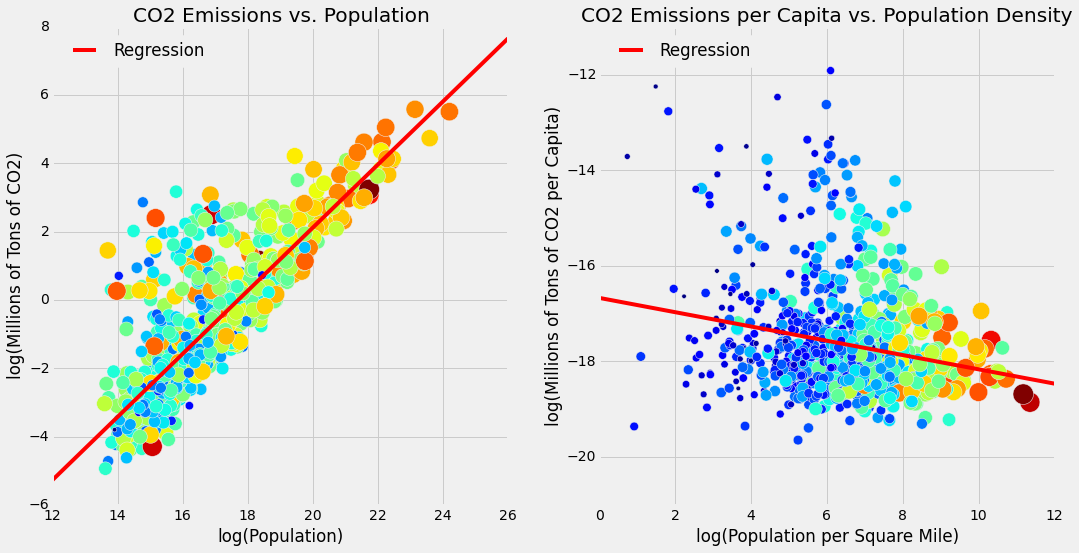


Figure 2: MSA level log(Total Emissions) vs. log(population) (left) and vs. log(population density) (right). In the left graph point size and color is based on the county’s land area with larger/redder points indicating larger area and in the right size and color is based on total population of the county with larger and redder points having a higher population. Red line shows the OLS regression. For the left graph the slope of the line is 0.9238 and for the right it’s -0.1496.

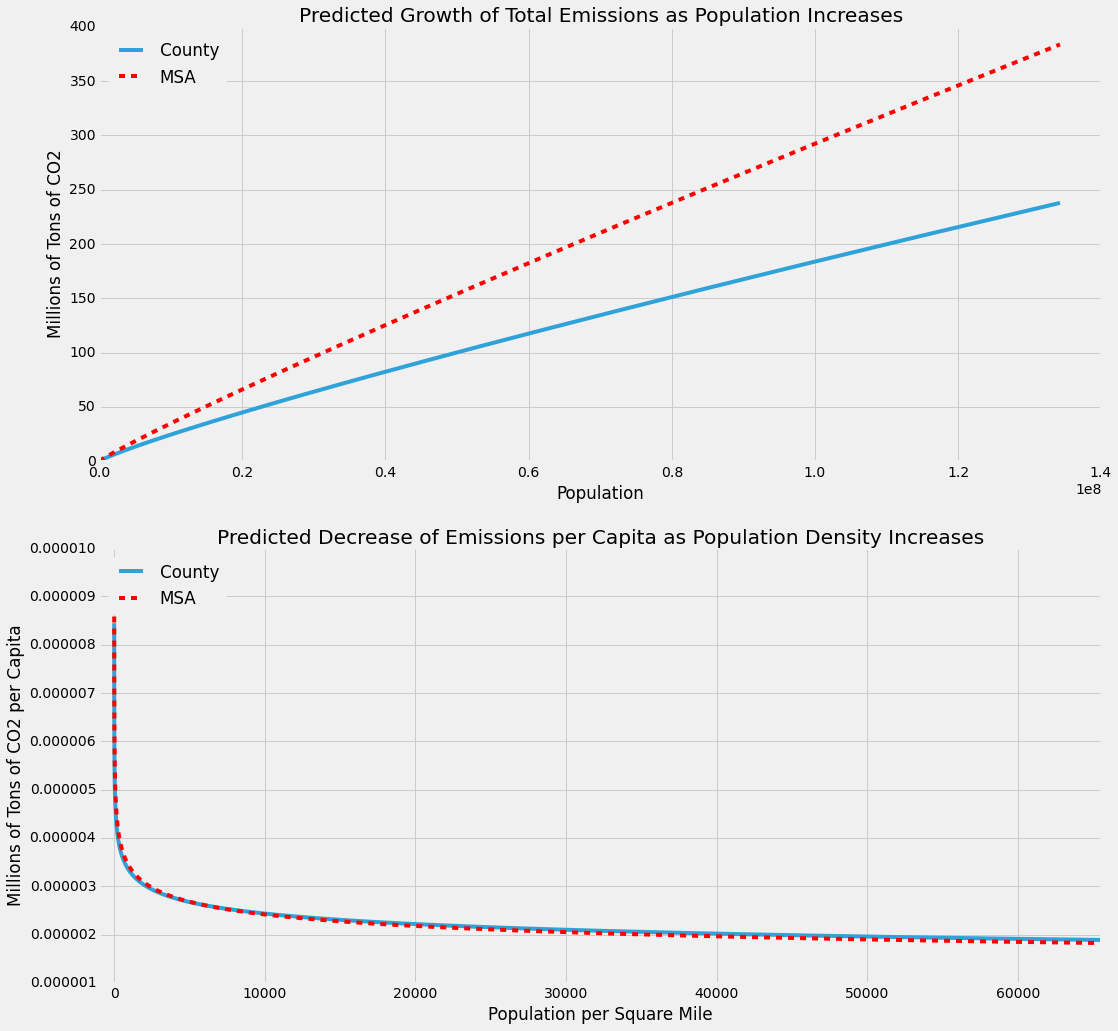


Figure 3: Results from County level and MSA level regressions plotted on a linear scale. As population increases total emissions grow slightly more less than linearly. As population density grows emissions per capita falls off rapidly until flattening out around 10000-20000 people per square mile.

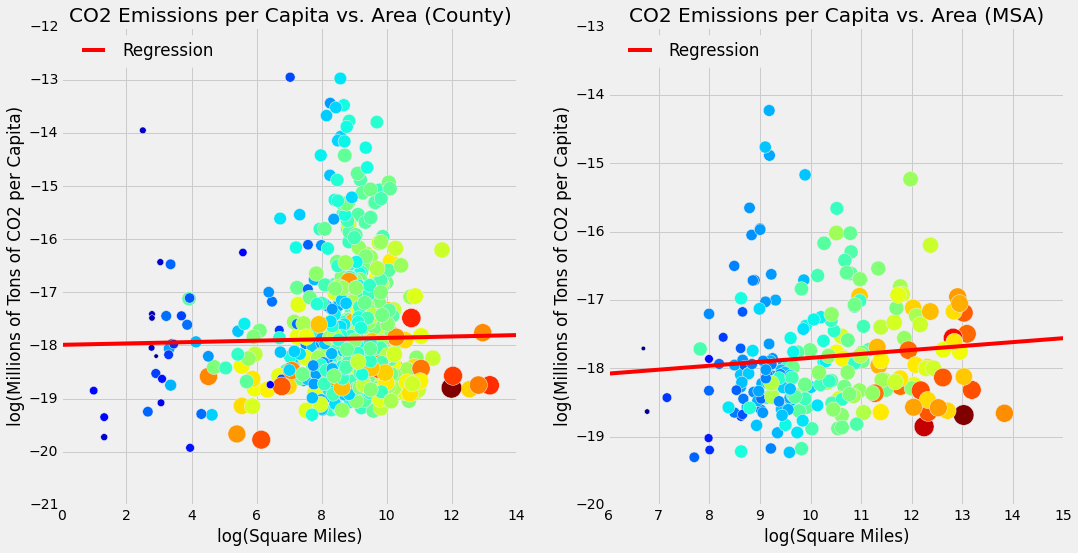


Figure 4: County level (left) and MSA level (right) log (Emissions per capita) vs. log(square miles). Points are colored and sized based on population. Red line shows the OLS regression. Both regression results include 0 in the 95% confidence interval.