Final Project Proposal

Project Title

Title: Air Pollution and Income, Population, Diseases, and Industry

Team members: Yu-Chun Chen A13356506, Yanyu Tao A13961185

Labor distribution:

Yu-Chen: Word analysis, exploratory

Yanyu: Background research, dataset,

Both: Question & hypothesis, data cleaning, coding

Questions Proposed

We are interested in the research about how air pollution, such as Toxic Release Inventory (TRI) and Greenhouse Gas Reporting Inventory (GHG), in the United States is associated with local geography. Specifically, we would like to explore how each county's business or factory types, average income, population, diseases, area of land/water are related to the seriousness of air pollution in its range. We expect that areas with more serious air pollution are those with more factories associated to lead, mining, and chemicals will produce more air pollutants, lower average income, higher population, higher incidence rate of respiratory or lung diseases. The issue is at stake as air pollution can damage not only human health, but also animals' and plants'. As most air pollutants are created by human activities, such as burning fossil fuels and emissions of factories, we would like to stress the danger of air pollution to our community so that we can promote a healthier place to live.

We can apply our analysis result to promote health. If we learn that a specific type of factories has a strong association with air pollution, this suggests that it is possibly a good idea to limit the number of this type of factories in each county. We can also distribute resources for hospitalization and emergency departments to each area using our analysis. We will focus our expected audience on local government and state governments, because they have the authority to regulate businesses regarding their emissions. With our analysis, governments will understand what types of areas produce the most air pollution, and are able to deal with pollution in those areas accordingly.

Background and literatures

According to the research "Relation between income, air pollution and mortality: a cohort study" done by Finkelstein, Jerrett, DeLuca, Finkelstein, Verma, Chapman and Sears with their investigation on data for cities in Canada and the United States, the authors conclude that pollutant level is higher in areas that have lower income. Moreover, areas with higher pollutant level, such as sulfur dioxide, have higher mortality risk. As the strong association between air pollution and increasing morbidity and mortality being confirmed by more and more authoritative institutions and organizations, authors of the article "Fine-Particulate Air Pollution and Life Expectancy in the United Sates published by The New England Journal of Medicine", claim that sustaining reductions of pollution exposure will lead to a better life expectancy. The data on life expectancy, socioeconomic status, and demographic characteristics for U.S. counties during late 1970s to early 1980s, and late 1990s to early 2000s were developed. Multiple regression models were built to demonstrate the association between pollution reductions and shifts in life expectancy. The result

suggests that the increase of overall life expectancy is highly sensitive to the reductions of air pollution alone and not as sensitive to other features.

We expect to confirm both research results since we hypothesize that serious air pollution will have a series of effects on commuter health, income, and other social factors, which eventually leads to lower life expectancy.

References:

Finkelstein, M. M., Jerrett, M., DeLuca, P., Finkelstein, N., Verma, D. K., Chapman, K., & Sears, M. R. (2003, September 02). Relation between income, air pollution and mortality: A cohort study. Retrieved May 12, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC183288/

Pope, C. A., Ezzati, M., & Dockery, D. W. (2009). Fine-Particulate Air Pollution and Life Expectancy in the United States. New England Journal of Medicine, 360(4), 376-386. doi:10.1056/nejmsa0805646. Retrieved May 12, 2019, from https://www.nejm.org/doi/full/10.1056/NEJMsa0805646

Data sources

US Household Income Statistics:

The dataset includes information of mean/median household income, number of households, geographic location information such as state, county and city for each city recorded, accessed from https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations.

US Facility-Level Air Pollution (2010-2014):

The dataset includes information of facilities, their geographic information, industry type, NAICS code, and emissions, accessed from https://www.kaggle.com/jaseibert/us-facilitylevel-air-pollution-20102014.

U.S. Pollution Data:

The dataset includes information of air pollution in each city in the United States, specifically CO and SO2 emission level, accessed from https://www.kaggle.com/sogun3/uspollution.

Plan of analysis

First, we will clean all the datasets so that we only keep the latest records (the newest year) and drop unnecessary columns. Then, we will change the column names so that each only contains at most 10 characters. We will also replace NaN values in integer/float columns with 0 and NaN values in string columns with string 'NaN'. Then, after we convert all three datasets to SEDF, we will merge US Household Income Statistics df with U.S. Pollution Data df to get a SEDF that has information of income, number of households, pollutant levels for each city.

After we are done with the preprocessing steps, we would like to make use of the ArcGIS module so that we can create multiple layers, along with functions such as aggregate_points and find_existing_locations, and also benefit from the use of Geoenrichment. We will create a point layer for US Facility-Level Air Pollution, in which each point on the layer represents each facility, and also a polygon layer for United States counties. Then, we will aggregate city points in each county so that we have average income, total number of households and total population, average pollutant levels for each county. (We are also expecting to find a dataset or geoenrichment data so we have incidence rate for lung and respiratory diseases in each county). We can also use

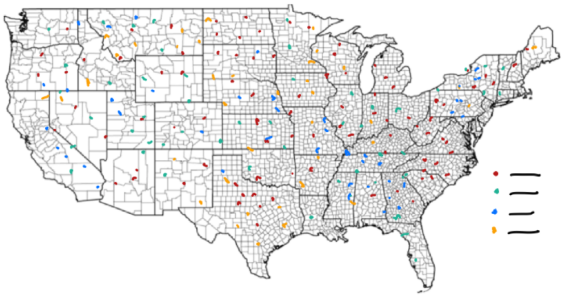
find_existing_locations to find out the number of each type of businesses/factories in each county. After doing all of these, we are able to create color-classed map for each county to show their income, population, incidence rate of diseases, and pollutant levels. We can also create a map, using both the county polygon layer and facilities point layer, with each facility colored or symbolized according to its industry type, to show how different type of business/factories distribute in each county.

Lastly, we will use Python sklearn libraries to perform regression and classification analysis. We would like to regress pollutant levels on population, income, incidence rate of disease, area to see how these factors are associated with air pollution. We will also divide each quantitative features to different levels of quantity range, and build a classifier to see how level of income, level of density, level of incidence rate of diseases, along with the most business type in each county, can help us classify the level of air pollution.

Layer 1: U.S. County

Layer 2: Facilities grouped by type

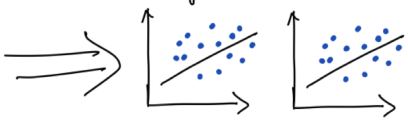
add layers



Color clustering
(Clustering income, population, etc. using cmap)

E.g. CA's counties

Regression models



Some issue: We are still wondering whether we want to use each air pollutant level, or Air Quality Index, or both to represent air pollution in our analysis.