---

1 Impacts of Toxic Pollution on Life Expectancy and Cancer Rates for Local Communities

Name and NetID of each member: Cole Smith (css459)

Member responsible for uploading submissions: Cole Smith

2 Background

2.1 Description

Using publicly available data from the EPA and CDC, I would like to explore the potential links of toxic dumping to general life expectancy and cancer rates at the local level, throughout the United States. Specifically, I have the data to do so from 2000 to 2016. Initially, I will constrain this range to 2010 to 2016 to align with the most recent Census (needed for life expectancy).

It is well known that certain chemicals are a contributing cause to a decline in public health. However, the process of disposing or sequestering these toxics are not equal, and local ecological factors may mean one method is significantly more dangerous for a given region than other methods.

2.2 Motivation

Environmental Accountability has historically been lower than required for long term ecological stability. Over time, the lack of accountability compounds to a large implicit cost for communities (and countries as a whole) as soil, air, and water supplies become more saturated with dangerous materials. The ability to identify a correlation between toxic dumping and public health is an important first step in fixing affected communities.

Therefore, my goal is to identify counties within the United States which have an above-average amount of industrial waste processing, and assess that area's life expectancy or cancer rates against national averages.

2.3 References

Cancer incidence and mortality around a hazardous waste depot
`https://www.ncbi.nlm.nih.gov/pubmed/9629820`

In this study, the authors found that, over an 18-year period, there was a significant rise in male brain cancer around a toxic dumping site. In addition, female breast cancer saw an increase on a shorter time scale. This study however was conducted in 1998. One possible outcome of my project could be to show that certain cancer rates have improved in recent times – An important confirmation of past efforts in environmental sustainability.

Hazardous waste and health impact: a systematic review of the scientific literature
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5637250/`

In this survey study, the authors assessed 57 papers on the health impacts of toxic waste. The overall results from many studies were inconclusive – There are many confounding factors which contribute to cancer rates among individuals. However, they found that oil industry waste to be the most hazardous of the types considered in the survey. This study is fairly recent, published in 2017, but again considers a traditional sample of selected communities, instead of a national survey.

3 Plans

3.1 Methodology

I have separated my data into three categories: Health (cancer and life expectancy), TRI (the Toxic Release Inventory), and the AQI (Air Quality Index). Given these three categories, I will determine some geo-spatial correlation between their values.

Data is grouped by various regions including state, county, and custom defined regions. I will use OpenStreetMap's Nominatim API to convert these regions to their Latitude and Longitude. State level information will be folded into county-level data points, and custom regions will be handled by changing the precision of the Lat/Lon measurement to a roughly 10km accuracy, grouping neighboring regions into one.

3.2 Proposed Experiments

It will first be important to treat this information as a small time series sample, at the yearly level. The deltas of observed values will be calculated from year-to-year from 2010 to 2016. This will be done for all possibilities of latitude and longitude, but the first step will be to assess the trends at a national level.

At the national and local level, a Time-Lagged Cross correlation between the three categories of signals can be found. I do not think there will be any strong correlations, so the next step is to form a regression problem using either Gradient Boosting or Elastic Net. With properly scaled features, the year itself can be treated as a variable.

3.3 Relevant Datasets

I will use the following datasets initially. Their discrete usefulness will be determined after a feature selection round with Gradient Boosting and Elastic Net.

TRI

Hospital Discharge Reason Data

Cancer (WONDER Dataset)

Daily AQI

Census Life Expectancy