Impacts of Toxic Pollution on Life Expectancy and Cancer Rates

Cole S. Smith
Courant Institute of Mathematical Sciences, New York University

Introduction

Using publicly available data from the EPA [3][2] and CDC [1], 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.

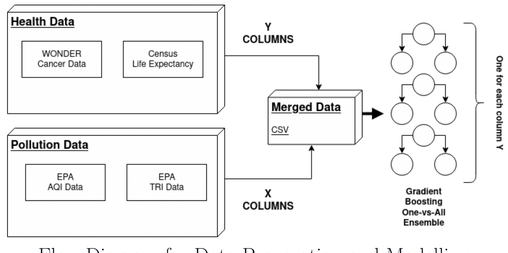
Motivation

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.

Methodology

Data Partitioning The data has been separated into three categories: *Health* (cancer and life expectancy), *TRI* (the Toxic Release Inventory), and the *AQI* (Air Quality Index). Given these three categories, I determine a 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 integer accuracy, grouping neighboring regions into one. This represents areas roughly 100km wide.

Predictive Modelling The columns I attempt to predict are defined in the Y set as the columns originating from the *Health* group of data. Namely, those are the discrete cancer rates, the overall age-adjusted rate, and the average life expectancy for a given area at a given year. These are then predicted via Gradient Boosting Regression and the R^2 value is assessed. There is one Gradient Boosting model for each column in Y, and these models are generated in a One-vs-All fashion against all columns in X – The Air Quality and Toxic dumping data from [2] [3].



Flow Diagram for Data Preparation and Modelling

Data Preparation

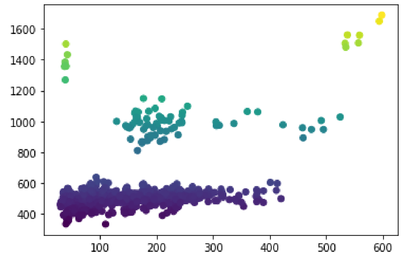
Merging Data The goal of the data preparation pipeline is to combine the Health, AQI, and TRI information into a single Pandas DataFrame. Since each of the data sets has its own unique schema, this meant transforming and scaling the data such that they all may be joined on a common key. Each row in the final data frame is a combination of **latitude, longitude, and year**. Overall, the final merged data set includes **586 rows** and **488 columns** (with EMA columns discussed in the next section) The variance of each column was evaluated, and there were found to be 4 columns of zero variance. Simple correlations were then calculated against the Age-Adjusted Rate of Cancer diagnoses:

Top 6 Correlations to Age-Adjusted Rate	
population	0.495891
defining: PM2.5	0.422032
aqi	0.419695
defining: Ozone	0.390418
defining: NO2	0.371733
sector: Petroleum Bulk Terminals	0.357192

Correlation Analysis From above, we can see that there are no columns in our design matrix that are immediately strongly correlated with the Age-Adjusted Rate of cancer levels for a given year and area.
This was expected however. Efforts to relate toxic dumping with cancer rates has been historically inconclusive as evidenced in [4].

Feature Extraction

Definition of Y Columns The Y columns, those for which we will label our design matrix and conduct our regression, are defined as the *Health* data sets as seen in the diagram in *Methodology*. Formally in the application, there are defined as columns with the following strings in their name: *age-adjusted_rate, count, cancer, population, life_expectancy.*
With these defined, we can now visualize potential clusters by pairing different X columns to Y columns.



Age-Adjusted Rate vs. AQI (Colored: Age-Adjusted Rate)

We can tell the relationship isn't linear, but there does exist defined clusters which group nicely with the age adjusted rate.

Exponential Moving Averages One such feature is the *exponential moving average* for each feature column. This is to model the time-weighted importance of each factor which may contribute to life expectancy and cancer rates, with the assumption that areas will consistently high levels of air pollution and toxic dumping exhibit higher rates and lower life expectancy than areas which have a higher recent spike in air pollution and toxic dumping. The EMA will capture this sentiment using one, two, and three year lags.

Algorithm

Training and Test Sets The data is split in a stratified way to prevent poisoning the test set. That is, the merged data set is *not unique by latitude and longitude* since it is considering multiple years for each lat,lon pair. Moreover, the values are *not on the same scale*, so all values are scaled to a normal range of $[-1, 1]$, preserving sparsity.

Model Selection Since we must consider nearly 500 columns in our data set, forward and backward feature selection with ElasticNet was found to be computationally infeasible due to the large search space. Instead, a model with selects features implicitly was chosen, **Gradient Boosting Regression**. This is a reasonable choice since it is a tree method which allows for non-linear relationships, automatically uses only the most relevant features, and returns to me the most important features for each and every Y column for which to predict.

Results

Performance In predicting the Age Adjusted Rate, and the frequency of specific cancers from AQI and TRI features the R^2 score was around **0.90 to 0.95** on average. To assess overfitting, this result was then confirmed by adding Gaussian random noise to the design matrix. The performance only dropped around 0.10 on average, confirming that the model was not simply memorizing data.

Findings More often than not, the most defining characteristics for different cancer rates were the number of *Petroleum Bulk Terminals* dumpings, and the *Air Quality Index* for a given area.
Surprisingly, the defining characteristic for the AQI that was most important in determining cancer rates was *PM2.5*. However, PM2.5 is often *not* a defining characteristic. Normally it is Ozone, which has been clinically proven to have an adverse affect on sensitive groups.
For "Age-Adjusted Rate" we see that "lat" and "lon" are both in the top 5 most important features for the GBM regressor which may indicate some of these relationships are spurious.

Future Work

Future work would need to be conducted in determining the causal relationships between these features and actual cancer rates. It is not enough to say that these correlations are linked. However, these studies are often expensive and **the novelty of this application is to show which factors are worth exploring.** For that, I believe **Air Quality** and **Tobacco, Petroleum Dumping** should be explored in greater detail to assess the health risk as it relates to life expectancy and cancers.

References

[1] Center for Disease Control. *WONDER Database 2010-2016*. data retrieved from CDC, `https://wonder.cdc.gov/controller/saved/D151/D71F287`. 2010-2016.

[2] Environmental Protection Agency. *Air Quality Index 2010-2016*. data retrieved from EPA, `https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw`. 2010-2016.

[3] Environmental Protection Agency. *Toxic Release Inventory 2010-2016*. data retrieved from EPA, `https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2017`. 2010-2016.

[4] L. Fazzo et al. "Hazardous waste and health impact: a systematic review of the scientific literature". In: *Environmental health : a global access science source* 16.1 (Oct. 2017). PMC5637250[pmcid], pp. 107–107. ISSN: 1476-069X. DOI: 10.1186/s12940-017-0311-8.