ENGM4676 – Final Project

Machine Learning in Public Health:
Predicting Regional Mortality Rate using Superfund Site Characteristics

Prepared by

Carter Hiltz (B00830210)
Aethan Cubitt (B00877256)

Submitted to

Dr. Issam Hammad
Dalhousie University
Halifax, Nova Scotia
12/03/2024

Table of Contents

List of Figures

List of Tables

Abstract

Since the Environmental Protection Agency (EPA) established the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) in 1980, it has managed the cleanup and containment of environmental disasters at “Superfund sites” throughout the country (US EPA, 2023). Many such sites exist, with 7% of Americans residing within one mile of one, and 23% within three miles (US EPA, 2023). Due to the variability in site severity, the diverse qualities of the hazardous material present, and the environmental implications inherent to each site, it is very challenging to quantify the health impacts on nearby populations. To address this issue, this study aims to employ machine learning and data analysis techniques to evaluate the impact of Superfund site proximity on public health outcomes. While previous studies have successfully identified a relationship between site proximity and composite public health indicators, this analysis intends to fill gaps in existing research by selecting a target that can reveal both the nature and severity of the impacts. To do so, county-level mortality statistics from The Institute for Health Metrics and Evaluation (IHME), categorized by various illnesses, diseases, and disorders, were analyzed in conjunction with data from the EPA pertaining to Superfund site location, affected area, and site severity score. Exploratory data analysis was applied to the county-level and state-level mortality rates, categorized by different causes, in relation to the presence of Superfund sites. Random Forest and Extreme Gradient Boosting machine learning regression models were developed across three different approaches to illustrate the relationship between mortality rate and the presence of Superfund sites. Unexpectedly, it was found that a marginal negative correlation existed between the mortality rate and Superfund site proximity, presumably due to the superior access to healthcare inherent to industrialized areas. The machine learning models struggled to attain reliable predictive performance when utilizing the selected features across the three different regression tasks. The study concluded that the presence of many confounding variables hindered the ability to isolate the impact of hazardous waste exposure on mortality rates. Future research would benefit from selecting a different target variable to quantify the nature and magnitude of the public health impacts attributed to Superfund sites.

Introduction

Pollution, contamination, and environmental degradation are common symptoms of industrialization. While precaution, procedure, and policy are effective deterrent strategies for accident prevention, a certain level of human error is inevitable. To account for such mistakes, the Environmental Protection Agency (EPA) created a program titled “Comprehensive Environmental Response, Compensation, and Liability Act” (CERCLA), commonly referred to as “Superfund”, tasked with restoration in the event of catastrophic environmental disasters (US EPA, 2019). Since it’s inception in 1980, the program has responded to 1881 highly contaminated sites throughout the USA, known as “Superfund sites” (US EPA, 2023). Of these locations, 1340 are still considered highly dangerous and remain quarantined to the public for further cleanup (US EPA, 2019). While the EPA makes every effort to remediate the environmental impact at these sites through the cleanup and containment of hazardous and toxic waste, the consequences of these catastrophes can only be mitigated, not entirely prevented. Because of this, it can be dangerous to live near a Superfund site, depending on the type of contamination or environmental disaster that occurred there.

Due to the substantial number of Superfund sites throughout the USA, many Americans live within close distance to one of these locations. A study by the EPA found that 7% of Americans live within 1 mile of a Superfund site, with 23% of Americans living within 3 miles (US EPA, 2023). While it is known that exposure to hazardous materials is detrimental to human health, the diverse health impacts of living nearby can be hard to quantify given the wide spectrum of contaminants present across all Superfund sites, the varying levels of severity from location to location, and the environmental implications of the area, like flooding or high winds. If an analysis was to be conducted to quantify these impacts, a target variable must be identified that is affected by all possible health outcomes resultant from Superfund site exposure.

Exposure to hazardous substances can cause a wide range of symptoms, from mild irritations to serious life-threatening conditions like cancer. It is universally understood that such substances harm human health and can consequentially shorten lifespan. Investigating this target variable was the ambition of a previous study, titled “The presence of Superfund sites as a determinant of life expectancy in the United States”, which found that populations living within a census tract of at least one superfund site could face a decrease in life expectancy by -0.186 ± 0.027 years (Kiaghadi et al., 2021). The success of this study, coupled with the absence of a comprehensive model to quantify the nature of the health impacts, and the relatively limited research on modeling geographical health trends, motivated investigation into other possible signals that could provide insight into their health influences on the American public.

It may be possible to identify a positive correlation between Superfund site proximity and mortality rate due to disease, cancer, illness, or other potential consequences of exposure. While life-expectancy is inherently linked to mortality rate, the latter is available with an added diagnostic marker in the form of “cause”, which could further illustrate the induced health consequences. Such data is publicly available on a per-county basis, courtesy of The Institute for Health Metrics and Evaluation (IHME) and could potentially provide enough granularity to conduct a conclusive study. The data pertaining to Superfund site locations, affected areas, and site severity is publicly available from the EPA as well, with the common variable being the location information. Through preprocessing and feature engineering, it would be possible to enrich the IHME dataset with the EPA Superfund data and perform subsequent analysis.

While it is undisputed that exposure to hazardous and toxic waste has a detrimental impact on health, the application of machine learning to this real-world problem could potentially reveal a relationship that could quantify the impacts of living near a Superfund cleanup site. This has extreme significance to the American public, as it could inform decisions on where to live, promote greater caution when handling toxic and hazardous materials, and influence environmental policy. Identifying that certain populations are at higher health risk due to their geographical distribution can potentially assist the EPA in their effort to mitigate the effects of these environmental catastrophes and provide additional resources to help coordination and resource allocation. Performing research in this area is of unquestionable significance, as it holds the potential to inform millions on the impact that their geographical distribution may be having on their health.

The objectives for this study are ambitious, as the statistical significance of the geographical distribution of environmental remediation zones is not yet known in relation to public health. The study objectives are as follows:

- Quantify the correlation between mortality rate and Superfund site characteristics
- Quantify the temporal impact of Superfund site designation over time on mortality rate
- Quantify the statistical difference in mortality data between counties with and without Superfund site designations.
- Develop machine learning regression models to perform the following supervised learning tasks and assess their performance:
    - Predict mortality rate variation over time by US county, given the number of Superfund sites by year, site severity, total affected area, and initial mortality rate.
    - Predict current mortality rate by US county, given the number of Superfund sites by year, site severity, total affected area, and initial mortality rate.
    - Predict the number of Superfund sites by US county, given the maximum, minimum, and average morality rate by year, and the percentage variation of mortality rate over time.

The expected outcome of conducting this study is to identify a statistical relationship between the characteristics of Superfund sites -such as location, affected area, and site severity - and their impact on public health, through the analysis of mortality rates by region, cause, and percentage variations over time. If the data allows, it is this study's ambition to identify key impact areas affected by exposure and to measure the extent of their effects.

Literature Review

The impact of Superfund sites on the environment and public health is not a new field of study. The Superfund program was established in response to a rising public concern about the impacts of toxic waste dumps. Since it's creation, the program has played an important role in both cleaning up hazardous sites and raising awareness about the long-term impacts on public health and the environment. As a result, research on the environmental and health impacts of these sites has been conducted. The following sources were reviewed prior to finalizing the proposal to inform the project team about the scope of the topic, the existing research landscape, and the gaps in public understanding of the impacts of these sites.

"The presence of Superfund sites as a determinant of life expectancy in the United States", by Amin Kiaghadi, Hamed S. Rifai, and Clint N. Dawson.

This 2021 study concerns itself with the impact Superfund site proximity on life-expectancy in surrounding communities. The study emphasizes that significant cleanup efforts were not conducted prior to the program's creation and as such may have a greater influence on life-expectancy than high-severity sites that were promptly decontaminated. The study confirmed that on average living near a Superfund cleanup site could be associated with a lower life expectancy (-0.186 ± 0.027 years), and that this decrease could be more pronounced in areas with higher sociodemographic disadvantages (up to -1.22 years). This was estimated to be the cause of a multitude of factors:

- Vulnerable communities have lesser access to healthcare resources
- Superfund sites situated in economically disadvantaged areas generally had weaker cleanup strategies.

To perform the study, 2018 statewide census data on race, ethnicity, income, education, and medical insurance was collected from the National Historical Geographic Information System (NHGIS) database. Superfund site location information was collected from the US EPA's public database. An Ordinary Least Squares (OLS) linear regression model was designed to measure the impact on life expectancy for those living near a Superfund site. A Random Forest machine learning model was then developed to corroborate the findings of the OLS model. The study considers the sites as independent and equivalent entities, disregarding their relative hazard level or potential to spread pollutants by air/water or other environmental factors. The study classifies data as "near" if there is one or more Superfund sites within the county census tract. Due to the EPA database's available features on Superfund locations, the study claims that there is no national study on the regional effect of Superfund sites with a greater granularity than by US county census tract.

A limitation found when researching this study was that the target, life expectancy, is a composite indicator in the context of exposure to hazardous waste. While this value does serve as a useful metric of portraying the dangers of exposure and living within close proximity of these sites, it is multifactorial in the sense that it comprises all contributors that influence mortality. It is an unquestionably helpful variable in measuring the severity of the detrimental health impacts associated with site exposure, but it does not portray exactly what they are. Because of this,

deaths due to unrelated causes are included in the analysis and introduce noise into the machine learning models performance.

“The Developmental Consequences of Superfund Sites”, by Claudia Perscio, David Figlio, and Jeffrey Roth

This 2019 study analyzes population data on children born in Florida from 1994 to 2002, focusing on the impact of Superfund cleanup sites on negative cognitive and health outcomes through exposure. These sites have been shown to be situated disproportionately closer to socioeconomically disadvantaged communities, and the study points out that this is an obscure mechanism through which poverty impacts public health. The study ultimately found that proximity to a Superfund site and the resulting exposure of unborn children to toxic waste compounded in long-term cognitive and behavioural impacts beyond the scope of traditional birth indicators (like Apgar score and birth weight).

The study utilizes Floridian birth certificate data from 1994 to 2002 and links it to public school records from 1996 to 2012 to gather demographical information. Like the previous study, Superfund site location information was sampled from the US EPA’s public database. While there was no machine learning performed in the study, the authors did use statistical methods such as bias-reduced linearization and two-stage least squares (2SLS) analysis to draw conclusions on features in the aforementioned data. The features used to quantify cognitive ability were standardized reading and math test scores and cognitive disability diagnosis. To quantify behavioural development, the study evaluated the likelihood of children failing a grade or reported involvement in punishable incidents at school.

One potential limitation of the research is establishing a causal relationship between these cognitive and behavioural consequences in children and the proximity of Superfund sites. Both of these may be influenced by underlying socioeconomic factors, meaning that they may simply be correlated under a common cause.

“The benefit of environmental improvement in the southeastern United States: Evidence from a simultaneous model of cancer mortality, toxic chemical releases and house values”, by Chau-Sa Ho and Diane Hite

This 2008 study analyzes the economic and environmental health risks associated with toxic chemical releases from waste sites and industrial facilities by measuring their impact on property value and cancer mortality rates to quantify the value of environmental cleanup. The study develops a county-level database across 9 US states, where data pertaining to known hazardous waste sites, Superfund sites, and cancer mortality rates are used as features to predict property cost via regression. While economic impact is the target for this study and not applicable to our research, its methodology concerning geographical analysis of Superfund sites and cancer mortality rates are of particular interest to us.

The study expresses that cancer mortality rates may be function of environmental quality and should therefore be include in the statistical analysis when predicting property costs. These features are utilized by a two-stage least squares statistical model to perform regression analysis.

In their literature review, the researchers cite other studies evaluating the temporal analysis of property costs in relation to Superfund sites and their designation dates, where the finding suggested there was a socio-economic transformation that often resulted in US counties following the listing of such a site on the EPA's National Priorities List (NPL). While socioeconomics are out of the scope of our analysis, there is likely to be a negative correlation between the socioeconomic status of a neighborhood and it's access to healthcare resources, which furthers our understanding of the project background.

Unsurprisingly, the study found that the quantity of Superfund sites and quantity of toxic releases per county are positively correlated with cancer mortality rates. Interestingly, the study found that there was a statistically insignificant positive correlation between property value and Superfund site proximity. This could potentially be explained by higher property values in industrialized areas, which are far more likely to suffer catastrophic environmental disasters that rural areas where property values are lower.

Conducting a literature review was important in evaluating gaps in existing research, as well as informing our decision of which statistical and machine learning tools to apply in our project. A shared limitation in the first two reviewed studies was the potential for threats to internal validity. In the first study, the impact of proximity to hazardous waste was quantified using life expectancy. While this composite indicator does an effective job at conveying the danger and broader health risks, it fails to isolate the exact cause by placing all emphasis on the effect. Stating that "smoking kills" is not nearly as effective as portraying the interim health consequences that may ultimately result in death. In considering the limitations of the second study, it was realized that socioeconomic factors can play as significant of a role in outcomes in cognitive ability, behaviour, and even life expectancy, as does the proximity of a Superfund site. Because of this, a target for a machine learning model must be selected that can inform as much about the nature of the induced public health consequences as it can quantify the magnitude of the impact to overall public health. The final study was illustrative of how geographical and temporal analysis can be helpful in establishing a statistical relationship between dataset features, useful information on developing a county-level database, and insight into geographical relationships between mortality rate and Superfund site and known hazardous waste releases.

Problem Formulation

While Superfund sites are universally understood as dangerous and harmful to human health, the environment, and even the economy, year after year they perpetually grow in numbers. Despite environmental remediation and mitigation efforts, most of the sites to receive Superfund site designation are still listed on the EPA's National Priorities List (NPL) (US EPA, 2019). While the affected area of these environmental catastrophes continues to grow, so does the public concern towards their induced long-term health consequences. However, a clear and universal understanding of their impacts remains misunderstood, as it can be statistically challenging to uncover faint relationships in complex data over the course of decades without the integrity of such a study succumbing to confounding variables, errors in data collection and processing, or feature bias. While it is essential to prohibit the public from accessing these dangerous areas, the quarantining of these sites leads to a societal misunderstanding of their dangerous nature. It is common to assume remediation efforts imply the situation has been

resolved and that it may once again be safe to live nearby, but the longevity of Superfund sites on the NPL list tell a different story.

The central issue that this study aims to address is the potential link between negative public health outcomes and the location, size, and severity of Superfund sites. It is known from relevant research that the nature of the exposure impacts are not conclusively correlated, and that an effective deterrent to complex and unpredictable health impacts would be a more robust stance on waste management and environmental policy (Fazzo et al., 2017). While exposure to hazardous material is known to be detrimental, a sufficient figure or model does not exist to convey the risk or to inform which populations are at greater risk than others.

Many Superfund studies have been conducted at the regional level to inform nearby populations about the specific environmental risks, a common example being the exposure to lead enriched dust from a site demolition (Khoury, 2003). While these studies outline important specifics on a site-by-site basis, there is lacking amount of generalized study that outlines the induced macroscopic health trends. By selecting a target that is universally impacted by the exposure-led health effects, it may be possible to identify a signal that enhances the comprehension of the issues involved. It is believed that despite government policy measures and remediation efforts, communities living near superfund sites may experience long term health consequences, including premature death.

The potential statistical relationship between Superfund site proximity and mortality rate is also significant for socioeconomic reasons, as demographical research has identified that disadvantaged and minority populations are at the highest risk of exposure (Persico et al., 2019). This has been described as a societal instrument through which poverty impacts public health, and can potentially perpetuate cycles of inequality through developmentally-induced health deficits (Persico et al., 2019). Identifying statistically significant relationships at a county level could guide policy, resource allocation, or influence public health interventions. Ultimately, it is believed that disclosing research findings on this subject, regardless of their magnitude, would contribute to a greater social equity.

The analysis should be centered on an informative target; a representative metric of health impact that is available on as large a scale as possible. While it could yield very interesting results to narrow in on a particular effect of site exposure, the aim of this study is to identify generalized insights. For this, mortality rate was chosen as the target, as it is available in sufficient quantity for nearly every county of the United States, courtesy of the Institute for Health Metrics and Evaluation (IHME). An immediate challenge faced during the acquisition of this data was that it was only available at on a per-county scale, whereas some studies had been able to secure health data on a census-tract level, offering superior granularity and resolution. US census data was not available to us, so a county-level granularity for our geographical analysis was chosen. While sufficient Superfund data is available from the EPA's database, the characteristics used in performing the analysis – such as location, affected area, and site severity - were available from in different data sources. Additionally, significant data preprocessing would have to be conducted to align this data into a format suitable for geographical analysis. This posed a risk of corruption due to data mishandling during the preprocessing and feature engineering stage, as these desired features would have to be assembled into a single data frame from numerous different sources.

There are broad and impactful applications for the findings that could result from research in this area. Public health agencies could use the results to advocate for stricter regulations and expedited remediation of sites. Policymakers could benefit from additional information when allocating federal resources to better address communities at higher exposure risk. Machine learning predictions could aid in the formulation of safety procedures in surrounding communities when a Superfund site is designated nearby and aid in informing the public of the associated risks and potential long-term impacts. Ultimately, this research aims to bridge the gap between environmental contamination and public health, driving actionable insights to protect vulnerable populations, and mitigate increased mortality risks associated with Superfund site exposure.

Data Collection, Processing, and Feature Engineering

Datasets were chosen to help define the relationship between Superfund site characteristics and population mortality rates at a county-level. An assumption made during the analysis was that the population within these counties and the number of Superfund sites would be generally equally distributed. The decision to make this assumption was driven by the following limitations and uncertainties:

- The highest degree of geographical granularity that could be found in mortality rate data was by county. While previous studies had been able to accomplish such a study with census-tract granularity, acquisition of this data requires an application approval.

- Environmental variables, such as high winds, rivers, or floodplains, are responsible for spreading contaminants from Superfund sites in varying degrees on a case-by-case basis. This impacts the statistical significance of direct proximity when making mortality predictions through regression.

To account for these factors that threaten the integrity of the analysis, it was decided that Superfund site proximity would be measured as “relative densities” by measuring the quantity of these sites on a per-county basis. While US counties are highly variable in size, it is also known that Superfund sites are generally situated near larger population centers in industrialized areas, as these are usually industrial accidents. Because of this, the assumption was made across each US county that the relative proportions of the population situated near such sites would be similar.

The datasets utilized in the analysis are listed, along with their relevant features and characteristics:

US County-level Mortality – The Institute of Health Metrics and Evaluation

IHME. (2014). *US county-level mortality*. Kaggle.com. https://www.kaggle.com/datasets/IHME/us-countylevel-mortality

While many datasets required preprocessing and feature engineering to prepare their features for the analysis, this one did not. This dataset was published to Kaggle.com by the IHME and itemizes mortality rate at the county-level and state-level. The dataset presents mortality rates for all 3143 US counties, across 21 all-encompassing causes (Neonatal disorders, etc). The data is available from 1980 to 2014, in five-year steps (with a final four-year step from 2010 to 2014). It also possesses a metric of percentage variation in mortality rate from 1980 – 2014 that is of great use to our analysis. In addition to providing average mortality rates across all these categories, it also presents maximum and minimum mortality rates. The county and state information is provided as a five digit Federal Information Processing System (FIPS) code.

The dataset features used in our analysis are listed below:

- **FIPS** - Five digit state-county code
- **Category** – All encompassing cause of mortality
- **Mortality Rate, X*** –Per capita, measured across regional population, max/min/average
    - **1980**
    - **1985**
    - **1990**
    - **1995**
    - **2000**
    - **2005**
    - **2010**
    - **2014**
- **% Change in Mortality Rate, 1980-2014** – Per capita, measured across regional population, max/min/average

All Current Superfund Sites – Environmental Protection Agency
(pdf contains .csv attachment)

Retrieved from:

United States Environmental Protection Agency (2024). *SEMS Superfund Public User Database.* Retrieved November 15, 2024, from https://semspub.epa.gov/work/HQ/401498.pdf

This dataset was made publicly available by the United States Environmental Protection Agency (EPA). It contains various geographical and identification information on all current Superfund sites. Of particular interest to this study is its Superfund site location data, which it provides in latitude/longitude coordinates. It also possesses NPL status designation dates for each of the listed sites, which is crucial for temporal analysis of the listed sites and their impact on regional mortality rates.

This data required some feature engineering before it could be utilized alongside the mortality data. To process this data into a form suitable for analysis, the site location coordinates were used to find the county FIPS code, which allowed comparison to the mortality data on a county-level. This was done by creating a Python function ("get_fips_code") to pass the coordinates in JSON strings to a web API, "Geocoder", managed by the United States Census Bureau, that returns FIPS codes (US Census Bureau, 2024). These FIPS codes were then used, alongside the designation date of each active Superfund site, to create new features in the mortality dataset pertaining to the number of active Superfund sites, by county, at each year where mortality rate data is provided (1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014). Since the IHME mortality dataset also possesses data at the state level, the number of Superfund sites was summed by state as well and appended to the mortality data frame.

The dataset features used in our analysis are listed below:

- **Latitude –** Superfund site coordinates
- **Longitude** – Superfund site coordinates
- **NPL Status Date** – Superfund site designation date

The features created from this dataset are as follows:

- **FIPS** – Superfund site location, by US state and county
- **Number of Sites in X*** – Summed quantity of Superfund sites by both state and county
    - 1980
    - 1985
    - 1990
    - 1995
    - 2000
    - 2005
    - 2010
    - 2014

EPA Superfund Site Details – Environmental Protection Agency

Retrieved from:

https://www.kaggle.com/datasets/thedevastator/epa-superfund-site-details

Data sourced from:

US EPA. (2018, September 13). *Search for Superfund Sites Where You Live | US EPA*. US EPA. https://www.epa.gov/superfund/search-superfund-sites-where-you-live

This dataset includes data collected by a Kaggle user “The Devastator” from a EPA database. It possesses various interesting features concerning Superfund sites, although the features of interest were “Site Score” and “Zip Code”. The site score is a metric used by the EPA to quantify the level of site severity, where the exposure risk is measured in ascending severity from 0 to 100. Feature engineering was conducted to concatenate this data into our existing mortality data frame using the identical “Zip Code” feature that is present in the “All Current Superfund Sites” data frame, which allowed the FIPS code to be added as a feature in this dataset. Using this feature, it was possible to attribute a “Site Score” feature in the associated rows of the mortality rate dataset, where 0 was inserted if a county did not possess any Superfund sites. If a county possessed more than a single Superfund site, the site score was summed to allow for the relative severity of each county to be compared. The site score was also summed at the state level, similar to the feature engineering performed in the “All Current Superfund Sites” dataset.

The dataset features used in our analysis are listed below:

Site Score – EPA standard metric of Superfund site severity, on a per-site basis

The features created from this dataset are as follows:

Site Score (Sum) – EPA site scores were summed according to how many sites were present in each county and state, to allow for relevant severity of exposure to be compared.

U.S. Federal Superfund Sites – Environmental Protection Agency

Retrieved from:

https://www.kaggle.com/datasets/srrobert50/federal-superfunds

Sourced from:

US EPA. (2015, August 14). *National Priorities List (NPL) Sites - by State | US EPA*. US EPA. https://www.epa.gov/superfund/national-priorities-list-npl-sites-state

This dataset includes data collected by a Kaggle user “4D4STRA” from a EPA database. It possesses a significant quantity of data pertaining to each census tract and mentions if they possess one or more Superfund sites via a feature “has_superfund”. While this dataset has the desired granularity to perform a study with high geographical resolution (at the census-tract level), there was not a mortality rate dataset found at the same granularity to compare it to. The primary feature of interest in this dataset was the metric of “LAND_AREA”, which provided a measure of the total area affected by Superfund site by location. Feature engineering was applied, using the “has_superfund” feature, to identify which “LAND_AREA” values should be added to a new column “LAND_AREA_SUM” in the mortality rate dataset. With these rows identified, it was a question of extracting the first five digits of the “FIPS_Block_Group” feature to produce a new “FIPS” feature and utilize this to sum “LAND_AREA” in the rows of the mortality dataset with identical “FIPS” values. The implementation of the

"SUM_LAND_AREA" feature into the final mortality rate data frame allows the area affected by hazardous waste, at a county and state-level, to be used as a feature in predicting mortality rate.

Data Analysis

Exploratory data analysis (EDA) was then used to investigate the nature of the statistical significance and relationships within the data. This will allow the structure, quality, correlations, and complexity of the chosen features to be established prior to the development of a model. Using the insights uncovered at this step, we can determine which regression models are most appropriate for predicting mortality rate, both instantaneous and over time, as well as the number of Superfund sites on a per-county basis.

The first stage in the EDA was to establish whether a statistically significant relationship existed between mortality rate and Superfund site proximity. To do so, the Pearson correlation coefficient was calculated between all mortality rate and number of superfund sites by county, from 1985 to 2014 in the time intervals mentioned. These Pearson correlation coefficients are plotted below in Figure 1:

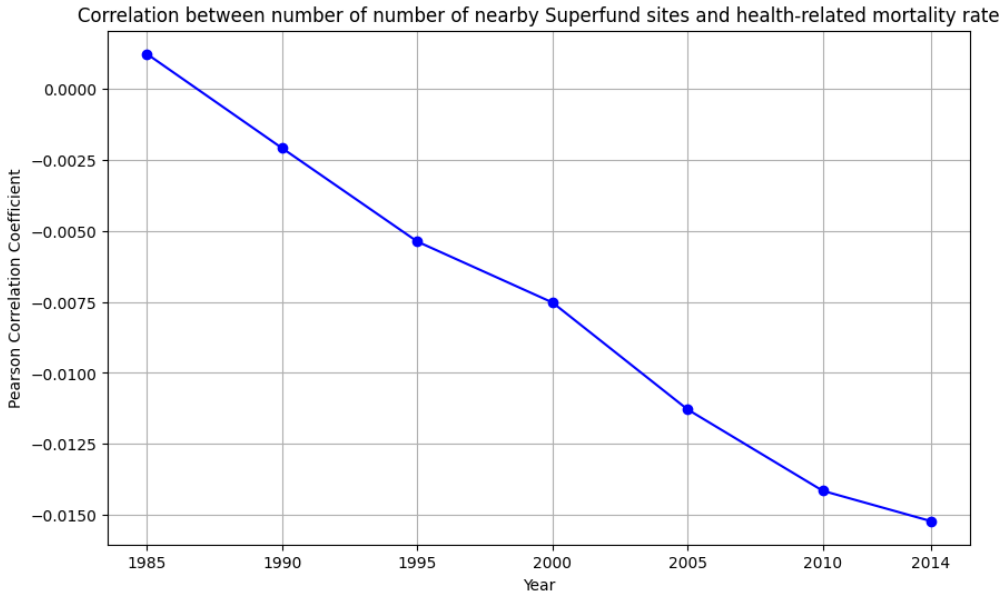

Figure 1: Correlation of Mortality Rate and Superfund Sites by County

Shown in Figure 1 above, the results of this analysis show that there is a steady negative correlation, growing in magnitude, between mortality rate and number of Superfund sites by US county. In other words, the data suggests that mortality rate is decreasing over time in counties that have larger numbers of Superfund sites. Although this result appears to contradict our

understanding of Superfund sites and their effects, we must consider the inherent characteristics of the feature. Superfund sites tend to cluster in industrialized areas. There areas would likely have superior access to healthcare than rural areas, and therefore the effect on mortality of living near such as site may be overshadowed by the benefits of greater healthcare access. To further analyze this, the correlation coefficient was computed across all 21 causes in the dataset. A sample of these correlation plot can be found below, with the complete set included in the Jupyter Notebook file.

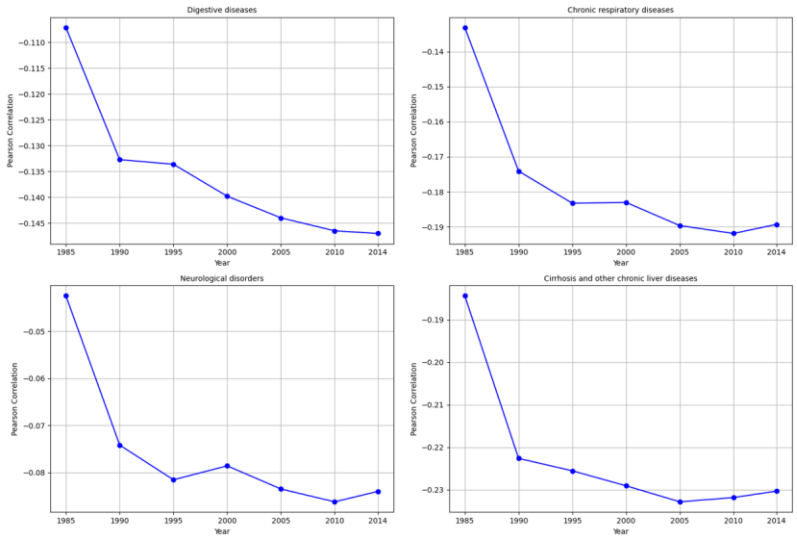

Figure 2: Correlation off Mortality Rate and Superfund sites by County and Cause

The consistency of this pattern across all reported diseases, illness, disorders led us to wonder if there is a reporting bias present in the data caused by the nearby Superfund sites. It is possible that a focus on other health issues is causing mortalities due to these causes to be underreported. Further analysis will be needed to identify if a surge or positive trend can be identified elsewhere. The exploratory data analysis is then elevated to the state level.

The correlation matrix shown in Figure 3 outlines the temporal analysis of the state-level correlation between mortality rate and number of Superfund sites, over the years listed.

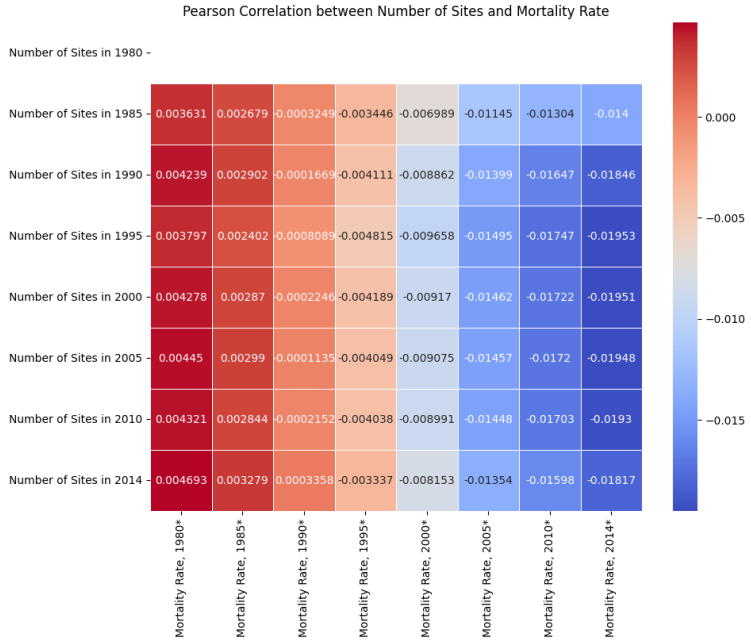

Figure 3: Correlation Matrix - Temporal Analysis of State Level Mortality

As can be seen from the correlation matrix, the correlation strength between mortality and number of nearby Superfund sites is very faint. While a relationship does seem to exist, it is very faint and complex. While the general correlation is low, the magnitude of the correlation between the current data for Superfund site quantity and the mortality rate (at the county level) increases with time. This is an interesting trend, as it matches the slow accumulation of Superfund sites over time. These sites are generally quarantined over many decades once found, and this contributes to a slowly increasing level of general exposure. It is possible that the Pearson correlation is not a suitable tool to quantify the relationship between mortality rate and number of superfund sites in each county. This could be due to the following factors:

- Non-linear relationship between mortality and superfund site proximity
- Confounding variables (healthcare access, population density, socioeconomic variables)

The Mann-Whitney U test is a suitable method of evaluating the effect of Superfund sites on mortality rates at a county level. It is designed to identify whether there is a tangible difference between two groups and is suitable for our application due to it's robustness to outliers and ability to compare the entire distribution of either group.

The results for the Mann-Whitney U test can be found in Table 1.

Table 1: Results of Mann-Whitney U Test on Mortality in counties with/without Superfund sites

Year	U Statistic	P-Value	Verdict ($p < 0.05$)
1985	138541423.5	0.0357066882	Statistically significant
1990	254862758.0	0.0030467154	Statistically significant
1995	274206454.5	0.0006960028	Statistically significant
2000	295481799.0	0.0017074547	Statistically significant
2005	308623300.5	0.0001093056	Statistically significant
2010	321953459.0	2.1007563e-05	Statistically significant
2014	335716733.0	2.7856389e-05	Statistically significant

From the Mann-Whitney test, there is a statistically significant difference in mortality rates for between counties that have superfund sites, and those without. A final correlation matrix was generated to further illustrate the nature of the inter-feature relationships. This final correlation matrix, where the data is analyzed on a county-level, is shown below in Figure 4.

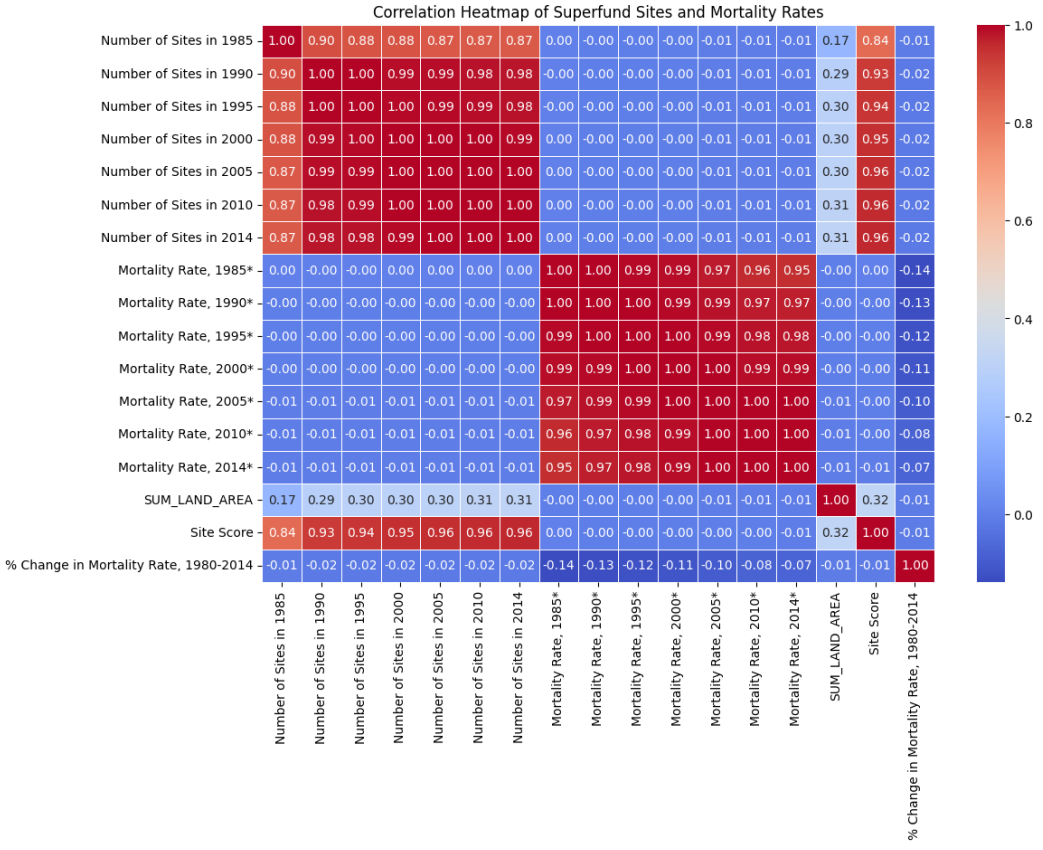

Figure 4: County-Level Intercorrelation of all Dataset Features

While the results of the intercorrelation between our features do not look very promising it is worth considering that the Pearson correlation coefficient quantifies the linear relationship that can exist between data. Should this relationship be logarithmic, it could be very faint and

hard to detect with such statistical tests. In the next stage of EDA, we will use scatter plot with logarithmic and square-root transformations to try and identify a relationship within the data.

While many iterations of the logarithmic and square-root transformation scatter plot testing were performed, the primary findings and plots are shown below. In a round of plot creation, a scatter plot was used with square-rooted superfund counts, as square root transformations are helpful in identifying relationships between discrete count data, such as the Superfund site count. Many such plots were generated, with an example shown in Figure 5.

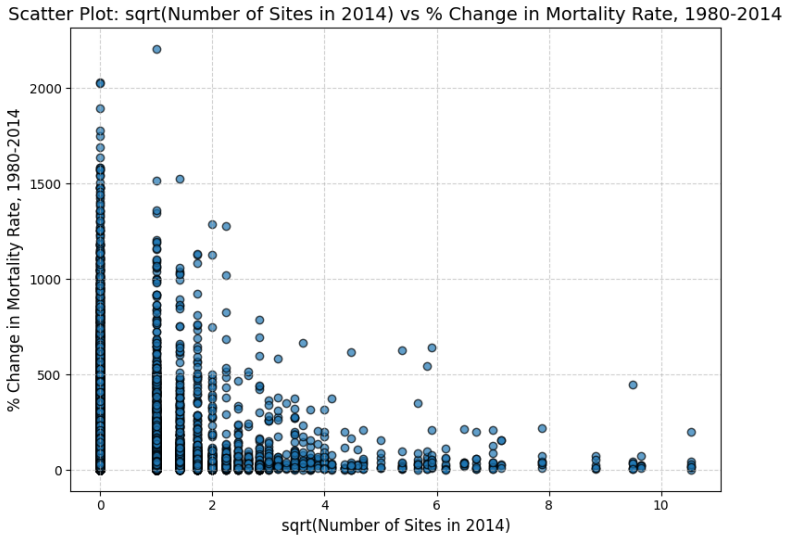

Figure 5: Square-Root Transformation Scatter Plot of Mortality Rate and Superfund Site Count

These plots suggest that a drop in mortality rate occurs as the number of nearby Superfund sites increase. Interestingly, the outliers do not reach as high as the count increases. While this does seem generally inconclusive, there is a faint trend that can be seen.

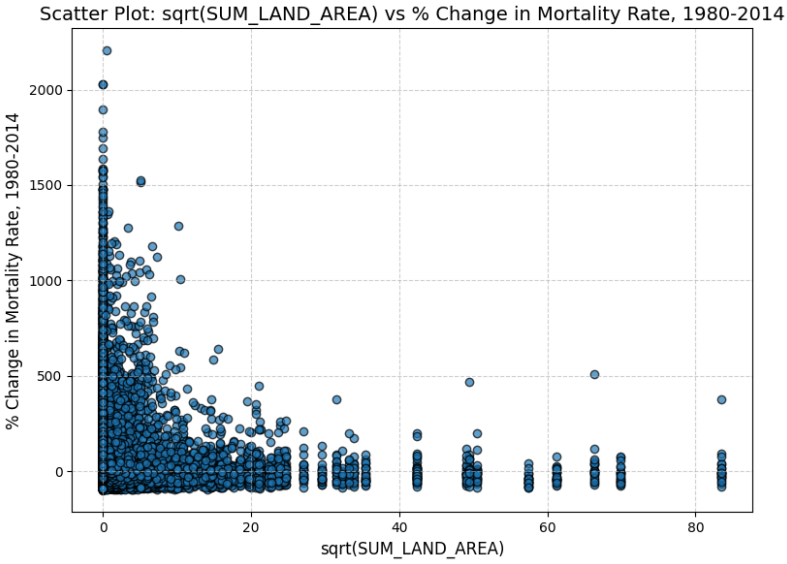

Figure 6: Square-Root Transformation Scatter Plot of Mortality Rate and Affected Area

Figure 6 depicts the square-root transformation of the affected land area against the percent variation in mortality rate over the entirety of the defined temporal range of the study. From the log-transformed scatter plots it was clear that sites with lower EPA severity score have a wider range of percentage changes in mortality rate. In the case of affected land area, there is a noticeable cluster at lower mortality rate variation and smaller land area, potentially suggesting there is a link between larger affected regions and a decline in public health.

Model Development

Finally, the results of the EDA can be used to make educated decisions on the types of machine learning regression models to be used for prediction. To restate the study objectives, machine learning models will be developed and utilized for the following three approaches:

- **Approach 1** - Predict mortality rate variation over time by US county, given the number of Superfund sites by year, site severity, total affected area, and initial mortality rate.

- **Approach 2 -** Predict current mortality rate by US county, given the number of Superfund sites by year, site severity, total affected area, and initial mortality rate.

- **Approach 3 -** Predict the number of Superfund sites by US county, given the maximum, minimum, and average morality rate by year, and the percentage variation of mortality rate over time.

Based on the findings in the exploratory data analysis, it will be difficult to make airtight predictions of either mortality rate or Superfund site count, due to the faint and complex inter-correlations in the features. From the correlation study, it is known that if a statistically significant relationship exists that it is nonlinear. Additionally, it is important to consider that some of the features are significantly more statistically significant than others, so a model must be selected that is robust against the implementation of insignificant features. Data interpretability is also a very important feature in the choice of model, as the feature importance when making a target prediction can be as indicative as the result given the nature of this study. With these considerations in mind, the following two models were chosen to perform regression across the three listed approaches, using supervised learning.

Random Forest Regression

Random forest regression utilizes a bagging technique when formulating target predictions. The decision trees branch out independently and formulate their own predictions. These predictions are then averaged to obtain a final prediction (Sruthi, 2021). Random forest models are known to perform best on larger datasets, which is suitable for our final mortality dataset as it possesses 67032 rows. These models are also known to excel when handling complex datasets, which is essential for our applications (Sruthi, 2021). Finally, a key driver in selecting this model for our analysis is it's known ability to effectively identify non-linear relationships in data (Sruthi, 2021). It is known from the EDA stage of the study that the strength

of the linear relationship in the data is very faint, so a logarithmic-capable model must be selected.

Extreme Gradient Boosting

Extreme gradient boosting models use a boosting technique when formulating target predictions, which involves new decision trees being added to the model at each iteration to correct the previous models' shortcomings. Like random forest models, it builds a model with multiple parallel decision trees. The extreme gradient boosting model is an optimized version of the gradient boosting model, with built in support for regularization, speed and performance improvements, and decreased memory usage (Analytics Vidhya, 2018). It is known to be highly efficient and capable of handling large datasets (Analytics Vidhya, 2018). It performs well with data where there is a clear relationship between the target and features, which may be an obstacle given the state of our data. However, it's inherent optimization and regularization features led to it's selection.

Approach 1 – Predicting Mortality Rate Variation over Time

To design the machine learning models to perform this prediction, an 80-20 train-test split was used, and "StandardScaler" was used across all feature data to preprocess the data through centering and scaling. In the case of the random forest regressor model, a standard 100 estimators was used at otherwise default settings. The extreme gradient boosting model was also initially configured to default settings, with 100 estimators.

The selected features were:

- **FIPS**
- **Number of sites in 1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014**
- **Mortality Rate in 1980***
- **SUM_LAND_AREA**
- **Site Score**

The target was **% Change in Mortality Rate, 1980 - 2014**

Following the first run, the performance of either model was improved through tuning of the hyperparameters to improve generalization.

In the case of the random forest model, it was very challenging to improve model performance when performing approach 1. This was done by defining minimum-maximum limits on each hyperparameter, such as n_estimators, max_depth, min_samples_split and others, and Randomized Search Cross Validation (using RandomCV) was used to generate 30 hyperparameter combinations of the hyperparameters, with 3-fold cross-validation applied to each combination. This marginally improved model performance.

In the case of the extreme gradient boosting model, performance improvements were attained in the first approach by increasing the number of estimators and perform more rounds of

boosting. The maximum number of tree splits was increased to a depth to 6, allowing more complex patterns to be identified in data. The alpha, beta, and gamma parameters were also tweaked to reduce overfitting and be more selective when "splitting" decision trees. An "Early Stopping" feature was also trialed and implemented into the XGB model at this stage, causing the iterations to cease if model performance does not improve after 10 rounds.

Approach 2 – Predicting Current Mortality Rate

Similar to the first approach, an 80-20 train-test split was used, and the data was centered and scaled using "StandardScaler" in preprocessing. Both the random forest regressor model and extreme gradient boosting model was also initially configured to default settings, with 100 estimators.

The selected features were:

- **FIPS**
- **Number of sites in 1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014**
- **Mortality Rate in 1980***
- **SUM_LAND_AREA**
- **Site Score**

The target was **Mortality Rate, 2014***

Following the first run, the performance of either model was improved through tuning of the hyperparameters to improve generalization.

For the random forest model, performance was improved by defining minimum-maximum limits on each hyperparameter and 3-fold, 30-iteration Randomized Search Cross Validation was used. Additionally, Extra Trees Regressor (ETR) was trialed at this stage to compare against random forest (RF). While ETR is similar to RF, ETR's splits are completely random when "growing" decision trees, and this randomness can sometimes offer performance improvements through a greater degree of regularization. These hyperparameter modifications only marginally improved the performance of the model against the test data.

The extreme gradient boosting model performance was improved by iteratively increasing the number of iterations and decreasing the learning rate to find a balance for maximum performance. While a cross validation solution could be devised to do this automatically – as had been previously done for random forest – there was a significant swing in performance that was observed when changing the hyperparameters, so they were tweaked manually to grow an intuitive understanding of their relevant effects. This was shown to considerably increase the model performance.

Approach 3 – Predicting the number of nearby Superfund sites using health data

An 80-20 train test split was used in the final approach, and the data was centered and scaled similar to the first two approaches. Default settings at 100 estimators were initially used for the first run of both the random forest and extreme gradient boosting models.

The selected features were:

- **Mortality Rate in 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2014**
  - **Average**
  - **Max**
  - **Min**
- **% Change in Mortality Rate, 1980-2014**
  - **Average**
  - **Max**
  - **Min**

The target was **Number of Sites in 2014**

Following the first run, the performance of either model was improved through tuning of the hyperparameters to improve generalization.

The random forest model performance was improved by increasing the number iterations from 100 to 150, and by decreasing the minimum number of samples from 2 to 4. This resulted in a marginal performance increase. Similar to the first two approaches, the random forest regressor model was relatively resistant to performance improvements and generalization. It was often found that the performance at the default settings was very close to its maximum observed performance after hyperparameter modification.

The extreme gradient boosting model performance was increased in this stage by drastically increasing the number of estimators from 100 to 500. Additionally, the learning rate was decreased to 0.25 from the default 0.3 to mitigate overfitting. The tree depth was increased from 6 to 8, and the alpha parameter was set to 0.1 to mitigating overfitting using lasso regularization. This resulted in a considerable performance increase.

Evaluation and Results

Following each execution of the regression analysis, performance metrics were recorded for each of the three specified approaches across both models. The performance indicators used at each stage are as follows:

- **Mean Squared Error**
- **R-squared**
- **Mean Absolute Error**

The following plots were generated at each stage to quantify model performance as well.

- **Feature Importance**
- **Predicted vs. Actual**

Approach 1 – Predicting Mortality Rate Variation over Time

Random Forest

First, random forest regression was utilized to predict mortality rate variation at a county level, given Superfund site locations and severity characteristics. It can be seen from Figure 7 that the model struggled to make accurate county-level predictions when using this data, even after significant regularization techniques were utilized. As can be seen from the feature importance plot, there is significant feature dominance from both the 1980 mortality rate data, and the FIPS location data. The 1980 mortality rate was included as a guide feature in this case, as it would be considerably challenging for the model to predict mortality variation over time without a "starting point". Otherwise, the "Number of Sites" Superfund count feature is surprisingly underutilized, informing us that the impact of their proximity is less indicative than we had previously thought. However, "Site Score" (severity measure) and "SUM_LAND_AREA" (relative affected-area measure) are sufficiently statistically significant when predicting county-level mortality variation. From the "Actual vs. Predicted" plot, the predictions are plagued by significant outliers heavily impacting the performance.

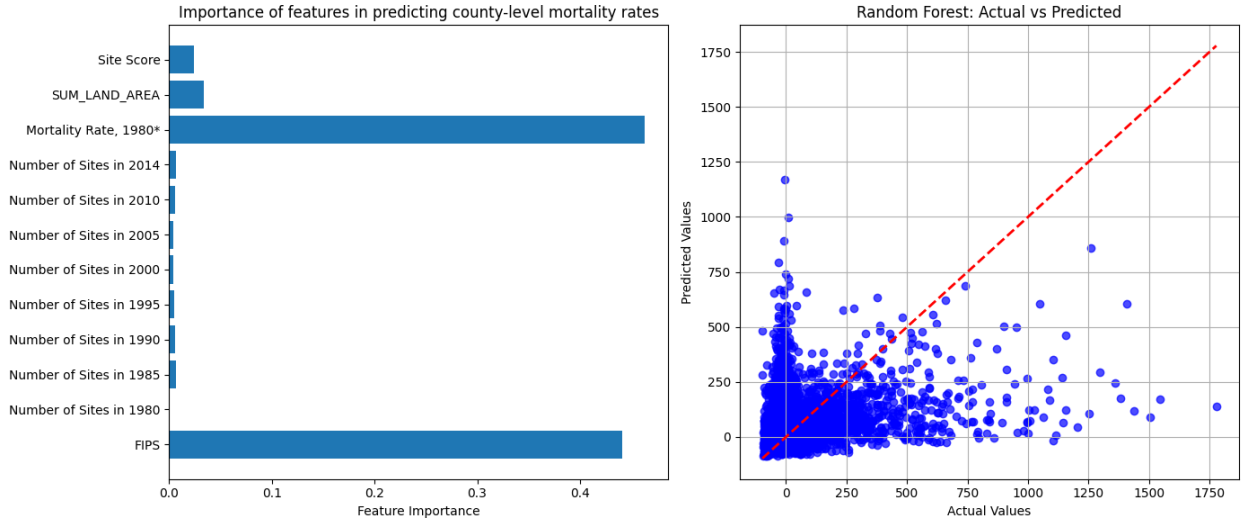

Figure 7: Approach 1 - Random Forest

The random forest regression performance metrics for approach 1 are listed below in Table 2.

Table 2: Approach 1 – Random Forest Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
10744.736	0.08731	47.2564

Extreme Gradient Boosting

Next, the extreme gradient boosting model was used to predict mortality rate variation at a county level as well. The feature importance and "Actual vs. Predicted" plots can be found below, illustrating the model's predictive performance. It is immediately clear that the feature importance has been distributed very differently when compared to the random forest performance, given the same data. The predictive importance of both "FIPS" and "Mortality Rate" still dominate the other features, but to a lesser magnitude than seen in the random forest model. In the case of the XGB model, the "Number of Sites in X" Superfund quantity feature is of much higher predictive importance, and this is reflected in a lower MSE and MAE than the random forest performance. Ultimately, the "Actual vs. Predicted" reveals that the predictive performance of this model can be greatly improved. Given the substantial effort to tune the model using the hyperparameters, it is believed that an inherent predictive limitation exists in the collected data.

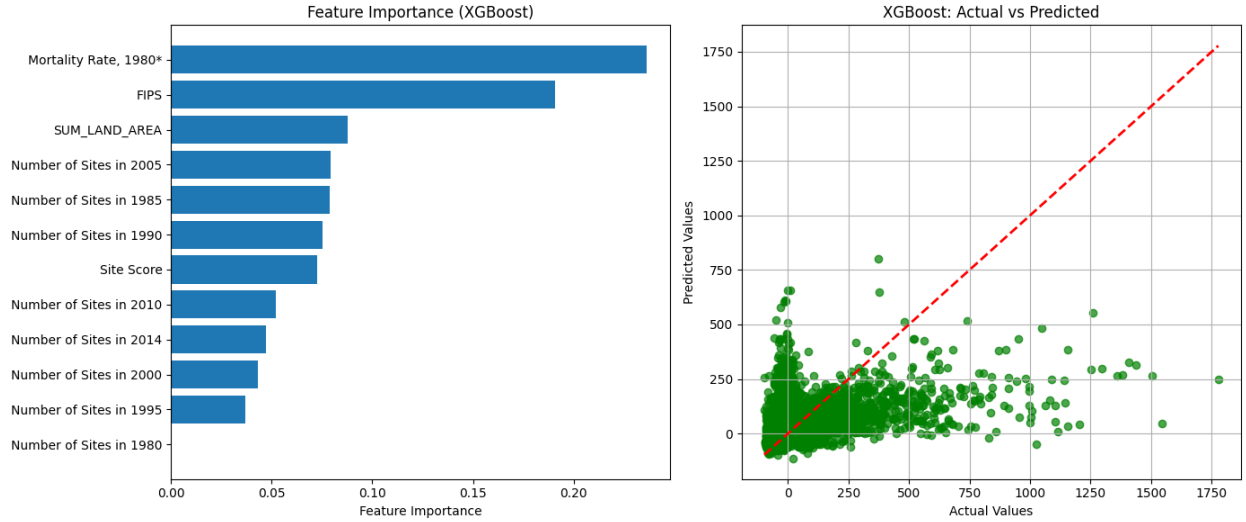

Figure 8: Approach 1 - Extreme Gradient Boosting

The extreme gradient boosting performance metrics for approach 1 are listed below in Table 3.

Table 3: Approach 1 - Extreme Gradient Boosting Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
9202.2358	0.218338	45.764449

Approach 2 – Predicting Current Mortality Rate

Random Forest

In the second approach, features pertaining to Superfund characteristics and the mortality rate in 1980 are used to predict the 2014 morality rate. Like approach 1, the mortality rate in 1980 was incorporated as a guide feature, intended to provide the model with an initial reference point. This feature, in conjunction with the Superfund characteristic data, was expected to assist the model in predicting future mortality rates. It is clear from the feature importance plot that this is not the case, and that the model has identified there is a low statistical significance between the number of nearby Superfund sites over the past few decades and the mortality rate, on a per-county basis. This is further confirmed by the passable performance of the model, where it can be seen from the "Actual vs. Predicted" plot that the model is clearly converging on the correct predictions.

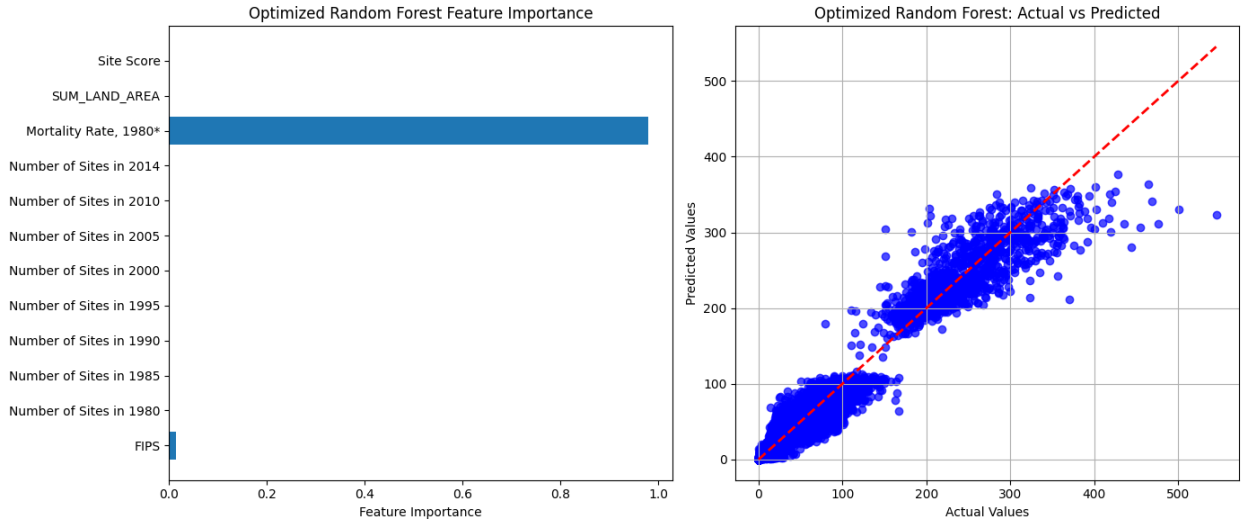

Figure 9: Approach 2 - Random Forest

The random forest regression performance metrics for approach 2 are listed below in Table 4.

Table 4: Approach 2 - Random Forest Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
189.915359	0.963299	7.339989

Extreme Gradient Boosting

The results of running the extreme gradient boosting model to predict the current mortality rate in approach 2 echo the findings of running the random forest model. The model is achieving passable performance by basing prediction almost entirely off the guide feature (Mortality Rate, 1980) without utilizing the county-level number of nearby Superfund sites to support its prediction. The analysis of the results from both models reveals that the correlative relationship between Superfund site proximity and public health outcomes is far more complex than initially anticipated. This complexity may pose significant challenges for a machine learning model to make accurate predictions without extensive preprocessing and feature engineering.

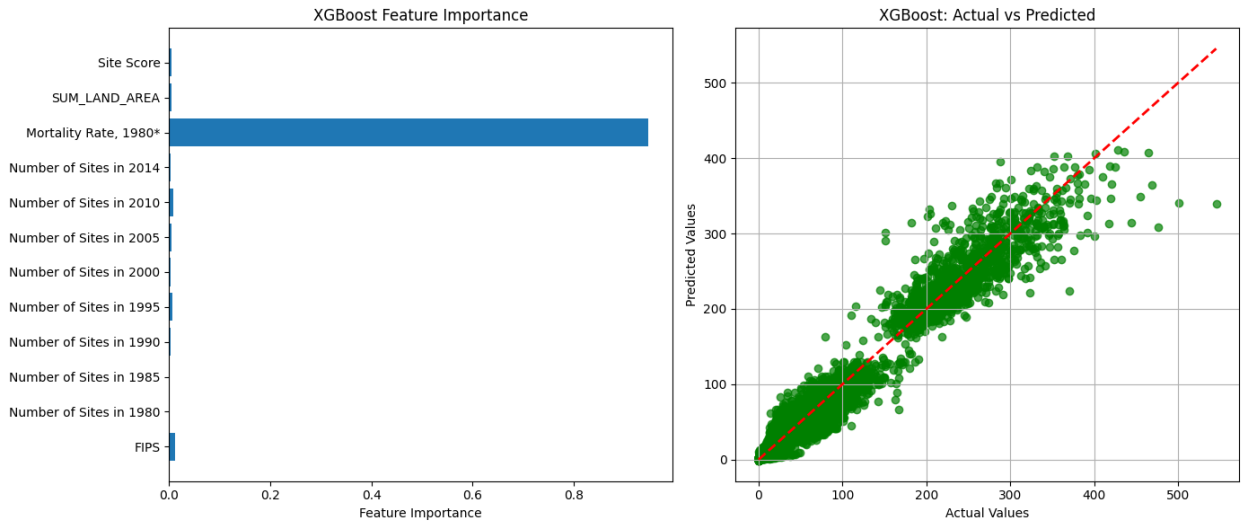

Figure 10: Approach 2 - Extreme Gradient Boosting

The extreme gradient boosting performance metrics for approach 2 are listed below in Table 5.

Table 5: Approach 2 - Extreme Gradient Boosting Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
187.20717	0.963822938	7.330596207

Approach 3 – Predicting the number of nearby Superfund sites using health data

Random Forest

For the third approach, the number of Superfund sites would be predicted per-county, utilizing nearly all mortality rate features provided in the IHME mortality dataset. In the case of the random forest model, the feature importance analysis reveals a well-distributed reliance on all features, with a particular emphasis on the variation in maximum, minimum, and average mortality rates over time. This is of great significance to the study, as it may not be feasible for a simpler and smaller model like RF or XGB to predict a complex value like the current mortality rate, but the model's reliance on mortality rate variation in this approach signals a tangible relationship between Superfund sites and the variation in county-wide mortality rates from 1980 to 2014. In terms of predictive performance, the model is unable to make reliable prediction, as can be seen from the "Actual vs. Predicted" plot.

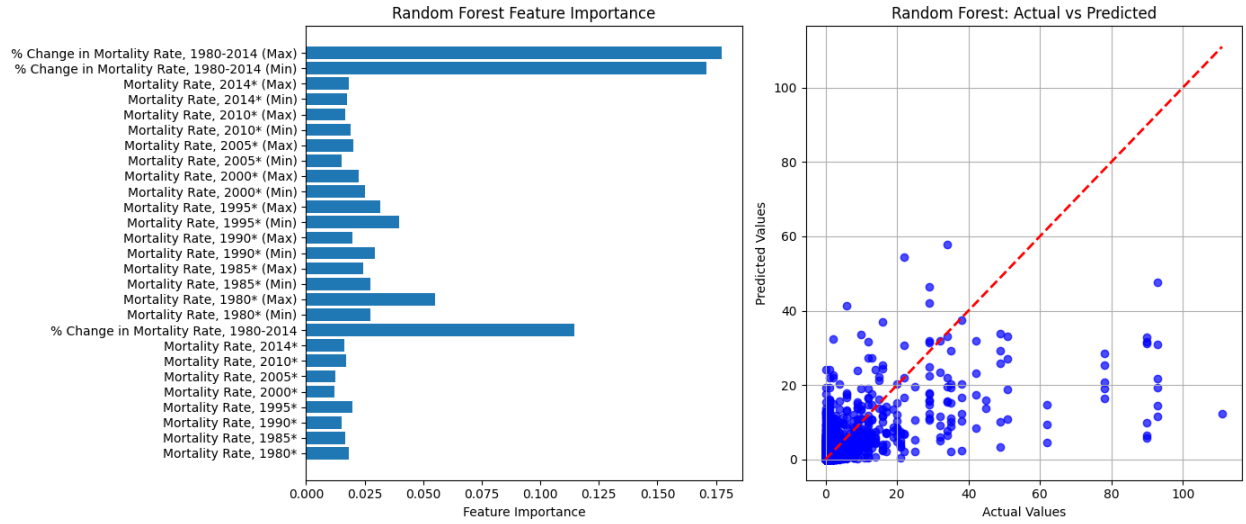

Figure 11: Approach 3 - Random Forest

The random forest regression performance metrics for approach 3 are listed below in Table 6.

Table 6: Approach 3 - Random Forest Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
11.94533657	0.4119755	0.82995264

Extreme Gradient Boosting

The XGB model aimed to predict the number of Superfund sites per county, using the same set of features provided to the random forest model. While it was observed that the extreme gradient boosting model tends to outperform the random forest model, in this approach it did not Similar to what was observed when utilizing the random forest model for approach 3, the model feature reliance is relatively equally distributed. While this gives the impression that the feature data is of high predictive quality, the unreliable performance of the model invalids this possibility. Ultimately, this instance is relatively inconclusive, as there is not a dominant feature or significant performance increase to draw conclusions from.

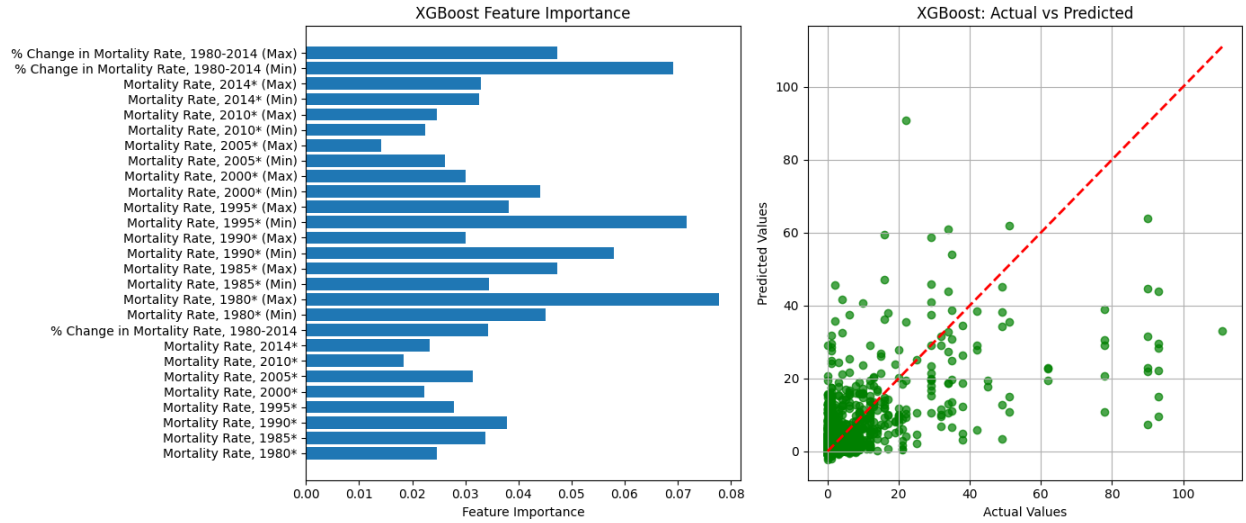

Figure 12: Approach 3 - Extreme Gradient Boosting

The extreme gradient boosting performance metrics for approach 1 are listed below in Table 7.

Table 7: Approach 3 - Extreme Gradient Boosting Performance

Mean Squared Error (MSE)	R-squared	Mean Absolute Error (MAE)
11.6430	0.42685759	0.849444997

Conclusion and Future Work

In conclusion, this study explored the complex relationship between the presence of Superfund sites throughout the United States, and their impact on public health. To perform this analysis, the mortality rate in the United States was studied on a county and state level and used as an indicative target to portray impacts on public health. Data pertaining to Superfund sites, such as location, severity, and area, was preprocessed into various new features to be utilized in numerous correlative and Mann-Whitney U testing performed during the exploratory data analysis and machine learning model testing and development. These analytical stages of the study were conducted with the goal and expected outcome of uncovering statistical insights concerning the strength and nature of a relationship between the presence of Superfund sites and regional mortality rate.

The scope of this study was ambitious, as previous studies under the same topic were able to demonstrate statistical relationships between public health and Superfund site location, using data with greater geographical granularity than we had access to. Additionally, nearly all the features in the mortality rate prediction stages had to be created using feature engineering from data sourced from multiple Superfund datasets, increasing the uncertainty of data corruption of mishandling. The results of the exploratory data analysis were mixed, but suggested a complex relationship existed between Superfund site presence and mortality. It was discovered that a faint negative correlation existed, and grew in magnitude year after year, between the number of Superfund sites and mortality. This was concluded to be likely caused by the tendency for such sites to be situated in industrialized areas, which overall would have better access to healthcare services and resources. It was also identified form the Mann-Whitney U test that county-level mortality data was significantly different than in counties that had such sites, and those that didn't.

From there, machine learning regression models were developed to perform supervised learning across three unique approaches to try and evaluate the statistical significance of the Superfund and mortality data, and the relative feature importance. These approaches involved predicting current mortality rate, mortality rate deviation over time, and number of Superfund sites, by county. These models were iteratively tested and hyperparameters were modified to increase their performance. While approach 1 identified that characteristics such as site severity and site area could hold statistical significance in predicting mortality rate on a county level, approach 3 identified that the mortality rate variation from 1980-2014 could hold statistical significance in predicting the number of nearby Superfund sites. The results of approach 2 were largely inconclusive. While the analysis provided insight into the relationship between mortality rate and Superfund site presence, there existed many factors that hindered the representative value of our results, such as insufficient geographical granularity, potentially confounding variables, and potential data mismanagement due to heavy feature engineering. A strong recommendation for future study in this field would be to analyze the Superfund data against other metrics of public health to develop a more all-encompassing and account for potential reporting bias. Finally, future expansions in this field of study would be benefit for developing larger deep learning models to process the data, as it became clear in the machine learning model development stage that the smaller random forest and extreme gradient boosting models would not be able to identify the complex relationship that may be present in the data to the extent that would be required.

References

Analytics Vidhya. (2018, September 6). *What is XGBoost Algorithm?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/#h-xgboost-model-benefits-and-attributes

Fazzo, L., Minichilli, F., Santoro, M., Ceccarini, A., Della Seta, M., Bianchi, F., Comba, P., & Martuzzi, M. (2017). Hazardous waste and health impact: a systematic review of the scientific literature. *Environmental Health*, *16*(1). https://doi.org/10.1186/s12940-017-0311-8

Ho, C.-S., & Hite, D. (2008). The benefit of environmental improvement in the southeastern United States: Evidence from a simultaneous model of cancer mortality, toxic chemical releases and house values. *Papers in Regional Science*, *87*(4), 589–604. https://doi.org/10.1111/j.1435-5957.2008.00179.x

IHME. (2014). *US county-level mortality*. Kaggle.com. https://www.kaggle.com/datasets/IHME/us-countylevel-mortality

Kiaghadi, A., Rifai, H. S., & Dawson, C. N. (2021). The presence of Superfund sites as a determinant of life expectancy in the United States. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-22249-2

Khoury, G. A., & Diamond, G. L. (2003). Risks to children from exposure to lead in air during remedial or removal activities at Superfund sites: A case study of the RSR lead smelter Superfund site†. *Journal of Exposure Science & Environmental Epidemiology*, *13*(1), 51–65. https://doi.org/10.1038/sj.jea.7500254

Persico, C., Figlio, D., & Roth, J. (2019). The Developmental Consequences of Superfund Sites. *Journal of Labor Economics*, *38*(4). https://doi.org/10.1086/706807

Sruthi, E. R. (2021, June 17). *Random Forest | Introduction to Random Forest Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

United States Environmental Protection Agency (2024). *SEMS Superfund Public User Database.* Retrieved November 15, 2024, from https://semspub.epa.gov/work/HQ/401498.pdf

US Census Bureau. (2024). *Welcome to Geocoder*. Geocoding.geo.census.gov. https://geocoding.geo.census.gov/geocoder/

US EPA. (2023). Population Surrounding 1,881 Superfund Sites [Review of Population Surrounding 1,881 Superfund Sites]. In https://www.epa.gov/. United States Environmental Protection Agency . https://www.epa.gov/system/files/documents/2023-08/FY22%20Population%20Estimates%20Superfund%20Final.pdf

US EPA. (2015, August 14). *National Priorities List (NPL) Sites - by State | US EPA*. US EPA. https://www.epa.gov/superfund/national-priorities-list-npl-sites-state

US EPA. (2018, September 13). *Search for Superfund Sites Where You Live | US EPA*. US EPA. https://www.epa.gov/superfund/search-superfund-sites-where-you-live

US EPA. (2019, September 13). Superfund | US EPA. US EPA. https://www.epa.gov/superfund/superfund-history

US EPA. (2019, April 23). *Superfund: National Priorities List (NPL) | US EPA*. US EPA. https://www.epa.gov/superfund/superfund-national-priorities-list-npl