ARE THERE ECONOMIC BENEFITS TO SUPERFUND SITE CLEANUP?

DS4A | GROUP 144

MEGAN WALKER
LEONARDO FALCON

EVANYA WILSON
NICOLAU ESTEVES

2021

AUGUST

TABLE OF CONTENTS

APPENDICES

All coding can be found in the appendices.

- Data Cleaning of All Variables
- Data Merging into a Single Dataset
- EDA
- Analysis

INTRODUCTION

Generally defined as contaminated sites due to hazardous waste being dumped, left out in the open, or otherwise improperly managed, superfund sites include, but are not limited to: manufacturing facilities, processing plants, landfills and mining sites. At these sites, high risk levels of environmental contaminants are identified which are known to have compounding adverse effects in the communities they surround. From polluting water, to air or land, these contaminants have the risk of negatively impacting not only the health and well-being of the surrounding communities, but economic growth as well.

In 1980, the United States Congress established the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), informally known as The Superfund. The fund was created to pay for the cleanup of the most hazardous waste sites in the country. Since the law was established, the worst sites have been brought under control, but according to the U.S. Environmental Protection Agency (EPA), there are still many sites exposing communities to dangerous levels of toxic chemicals. In fact, as of September 2020, 73 million Americans lived within 3 miles of a superfund site. This includes 23% of children under the age of five, 28% of all minorities, and 24% of all households below the national poverty level.

73 MIL

Americans live within three miles of a superfund site

RESEARCH QUESTIONS

For this project, we wanted to find out how has the 1980 establishment of the Superfund affected pollution in the surrounding areas? What have been the economic impacts for surrounding communities?

We expected to see an overall reduction in pollution; however, what was still unclear was how those reductions have affected some economic factors. We focused specifically on outcome variables, such as home ownership, and unemployment.

WHY DOES THIS PROBLEM MATTER?

The potential negative economic ramifications not only affect individuals directly impacted, but possibly generations to come. Furthermore, some of these ramifications may also hinder equal opportunity for health, wealth, and the pursuit of happiness since it affects disproportionately marginalized communities.

VARIABLES

The table below provides a summary of the data sources used while exploring the topic. In addition to the data link that leads directly to the source.

VARIABLE	SOURCE	CREDIBILITY
Superfund	EPA	Superfund Data and Reports	High	Data comes directly from the U.S. Environmental Protection Agency.
Air Quality	U.S. EPA. Air Data: Air Quality Data Collected at Outdoor Monitors Across the US	High	Data comes directly from monitoring locations from across the U.S. and is housed by the U.S. Environmental Protection Agency.
Water Quality	Water Quality Portal	High/Medium	Data comes directly from monitoring locations and are housed by the U.S. Environmental Protection Agency and U.S. Geological Services.
House Price Index	FHFA House Price Index	High	The FHFA House Price Index (FHFA HPI®) is the nation's only collection of public, freely available house price indexes that measure changes in single-family home values.
Unemployment Rate	Local Area Unemployment Statistics Home Page	High - Data comes from the U.S. Bureau of Labor Statistics

DATA CLEANING

The final database used in the project was a combination of multiple data sources, including air, water, unemployment, house prices and superfund sites across the US. The primary objective was to generate a unique database that provided a better opportunity for more efficient data exploration, analysis, modeling and visualization.

In general for each data source, all files were downloaded in csv format. After an initial data exploration and closer look into them , some columns were dropped from the dataset. Through some analysis resulting from the data feature process, new columns were added that proved to be more relevant in answering the project's main research question. All column names were revised to be lower case, using underscores, and clear descriptors. In addition, columns were updated with the appropriate data type.

Superfund Sites: The dataset contains the type of contaminant and whether human exposure exists, and if groundwater migration is under control. It also contains data of whether the required remedies have been met (Construction Complete variable) and the date that the remedy was completed (Construction Complete Date). This data set was used to create a focal point for our analysis. We looked at indicators (air quality, housing price, etc.) surrounding these sites. We compared indicators at the time the site was indicated as a superfund, the time it was cleaned, and in 5 year increments following the cleaning. The column headers were truncated since some were quite long or contained unallowable characters (e.g. " Groundwater Mitigation Under Control").

Air Quality: This data set provided index, statistics and monitored values for air quality in general and for specific types of air pollutants for all 50 states and its large cities across the U.S (mostly from 1980 to 2021). This was useful for collecting historical data for our geographic location of choice. Combined with unemployment, home price, and other economic factors, we got more insights on contaminants' negative economic impact. The air quality variable was created by adding the columns labeled "good" and "moderate", then dividing by the total number of days where air quality was measured in that county and year. This provided a percentage of good air quality days.

Water Quality: This dataset samples water quality at over 2.2 million sites across the U.S. The data include the total Nitrogen and/or Phosphorous levels in the water samples from 2009 to 2019 in a single csv file. The water quality variable was created by adding the number of Nitrogen and Phosphorous samples that exceeded the water quality criteria, then subtracting that from one and dividing it by the total number of usable samples collected. This provided a percentage of good water quality samples.

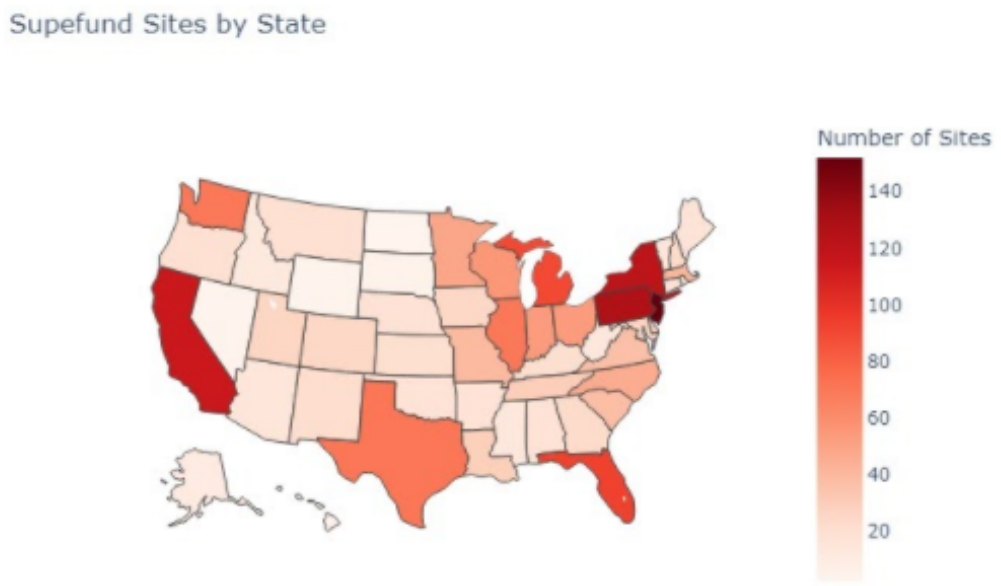
House Price Index: The data set used is called the Annual House Price Indexes (Five-Digit Zip Codes) provided by the Federal Housing Finance Agency. This data is not being collected on an ongoing basis, the historical data is already provided from the year 1991 to 2020. This dataset provided further insight into how Superfund sites impact the house market in nearby communities. Data was provided at national, census division, state, metro area, county, ZIP code, and census tract levels.

Unemployment: This dataset consisted of annual employment, unemployment, and labor force data for States and counties in the US from 1999 through 2019. The important variables considered for this dataset were State, County, Year and Unemployment Rate.

The combined data frame included over 49,000 rows and 41 columns. We then conducted the exploratory data analysis to help us narrow down our analysis.

EXPLORATORY DATA ANALYSIS

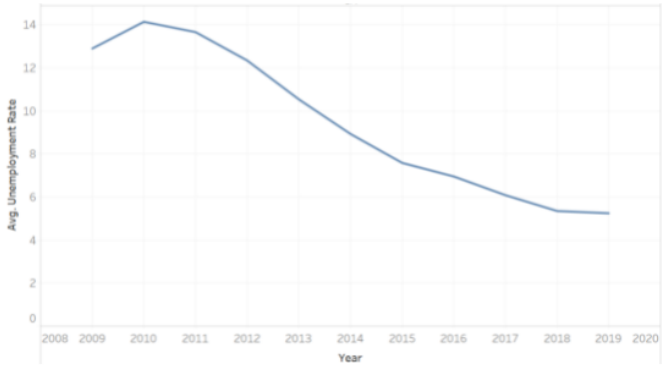
There had been a total of 1,834 superfund sites across the 50 States and the District of Columbia, and more than half of them were identified as having been cleaned.

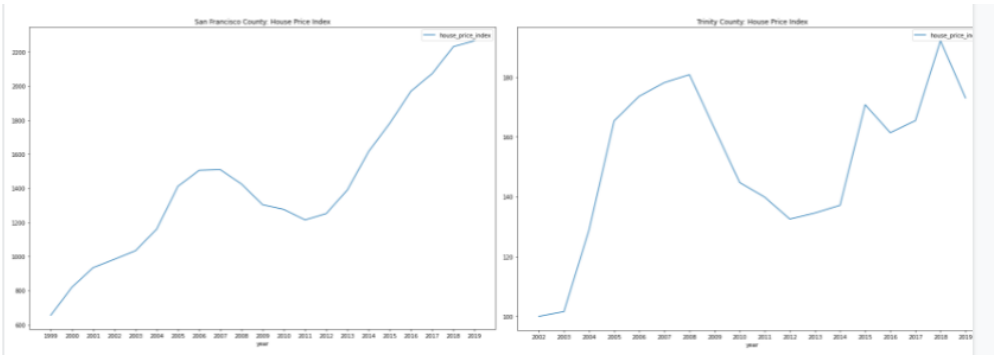


We further examined the data to identify any initial patterns that existed with regards to the sites, pollution, or the economic factors. Given the expansiveness of the dataset, we chose to identify one or two locations that could serve as the dataset case studies. In examining each domain, we found that California and New Jersey might be insightful to this end.

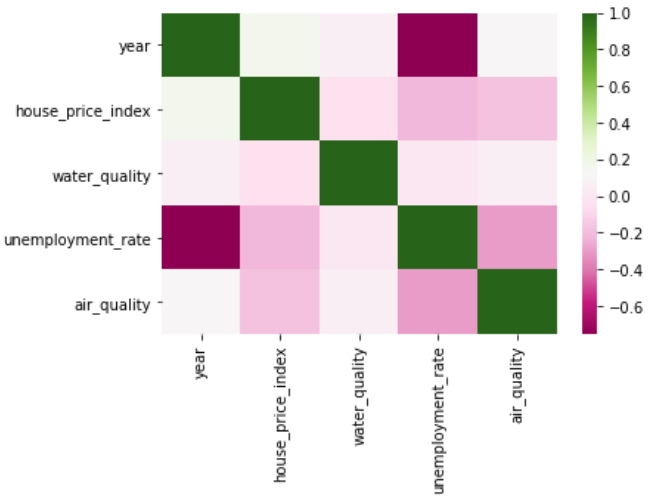
California and New Jersey had the greatest number of total superfund sites, yet were on opposite sides of the United States. In addition, California had the highest counts for other factors such as air pollution, water quality, unemployment, and house prices. This makes sense in part because of the geographic area and population size of the state.

Both California and New Jersey had some of the same pollution and economic patterns over time. Water and air pollution tended to decrease from 2009 to 2019, while house prices tended to increase significantly. In addition, we see in both states that unemployment decreases. These all seem to suggest economic and environmental successes in the ten year period.





We then began looking into the relationship between some pollution and economic factors. We found a consistent negative relationship between house prices and unemployment; as unemployment went down, house prices went up. There may have been a positive relationship between water quality and air quality, though this is less clear in the scatterplots.



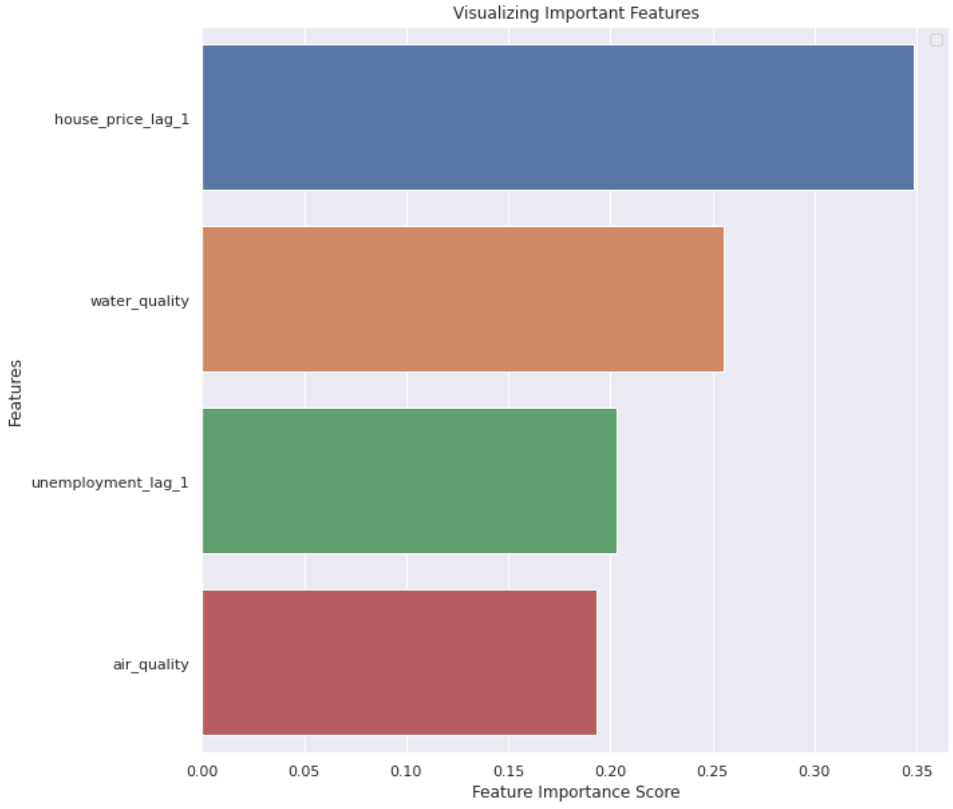
Using correlation heat maps, we were able to more clearly understand whether or not relationships existed between the variables. Narrowing our focus to a single county, San Bernardino in California, we continued to see the negative correlation between unemployment rate and house price index. There is also a negative correlation, albeit weak, between the house price index and water quality, specifically with regard to Nitrogen count.

Finally, we explored the differences that existed between clean and active sites. There appeared to be a statistically significant difference in employment rate among active vs. non-active sites in California and New Jersey. However, differences for other factors - air quality, water quality, and house price index - did not appear to be statistically significant. When we zoom out, however, and use the full dataset, we see statistically significant differences between clean and active sites. Clean sites seemed to have better air and water quality, higher house prices, and lower unemployment.

VARIABLE	CLEAN > NOT CLEAN	P-VALUE
Air Quality	Yes	<0.01
Water Quality	Yes	<0.01
House Price Index	Yes	<0.01
Unemployment Rate	No	<0.01

ANALYSIS & PREDICTIVE MODELING

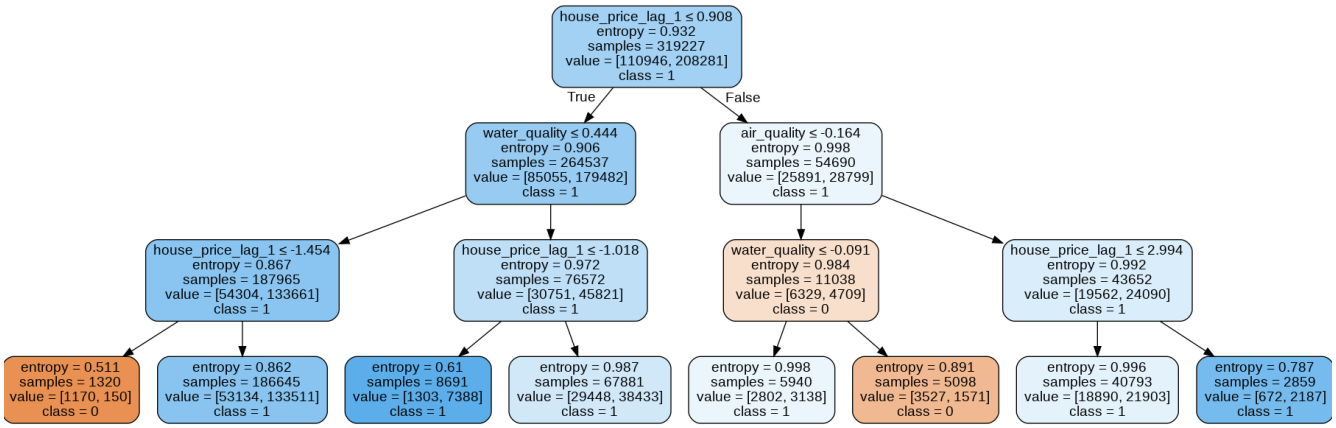
In our statistical analysis, we further explored the characteristics that differentiate clean sites from unclean sites. We did this by fitting our data into classification models.



We calculated the importance of the four features in the model: house price after one year, unemployment rate after one year, water quality, and air quality. We identified the house price after one year as having the greatest feature importance score at 0.34 and air quality with the least importance at 0.19. In order to ensure that the feature importance was not being skewed by the unit difference in the variables (percentages versus dollars), we used a Standard Scaler and conducted the feature importance calculation again. The results were the same with house price being identified as the most important feature.

However, when the features were pulled into the decision tree and random forest, the figure became unwieldy to get the highest level of accuracy.

We initially tested a few classification models to determine the ones that best fit our data. Of those we tested, the Random Forest and Decision tree models provided the highest accuracy with 86.0% and 75.5%, respectively. We then began a pruning technique by plotting the accuracy for the classification models given changes in depth. While the depth of 30 provided the greatest accuracy to the models, it becomes unwieldy and difficult to assign meaning. We decided to start with a depth of 5, which decreased the accuracy to 66%.



While we appreciated the information gleaned from the classification model, we ultimately determined that a regression model would best answer our original question of how superfund site clean up affects our economic factors. We built two models to this end: one with unemployment rate as the dependent factor and another with house price index as the dependent factor.

The results of those regression models are seen below. For each model, all variables were significant at 0.05. In the house price model, all variables seemed to have an inverse relationship with house price.

House Price Model

Dep. Variable:	house_price_lag_1	R-squared:	0.112
Model:	OLS	Adj. R-squared:	0.107

	coef	std err	t	P-value
Intercept	2230.0571	240.939	9.256	0.000
water_quality	-0.4768	0.495	-0.963	0.336
air_quality	-1735.1847	247.122	-7.022	0.000

In the unemployment model, air quality had an inverse relationship with unemployment, while water quality had a positive relationship with unemployment.

Unemployment Model

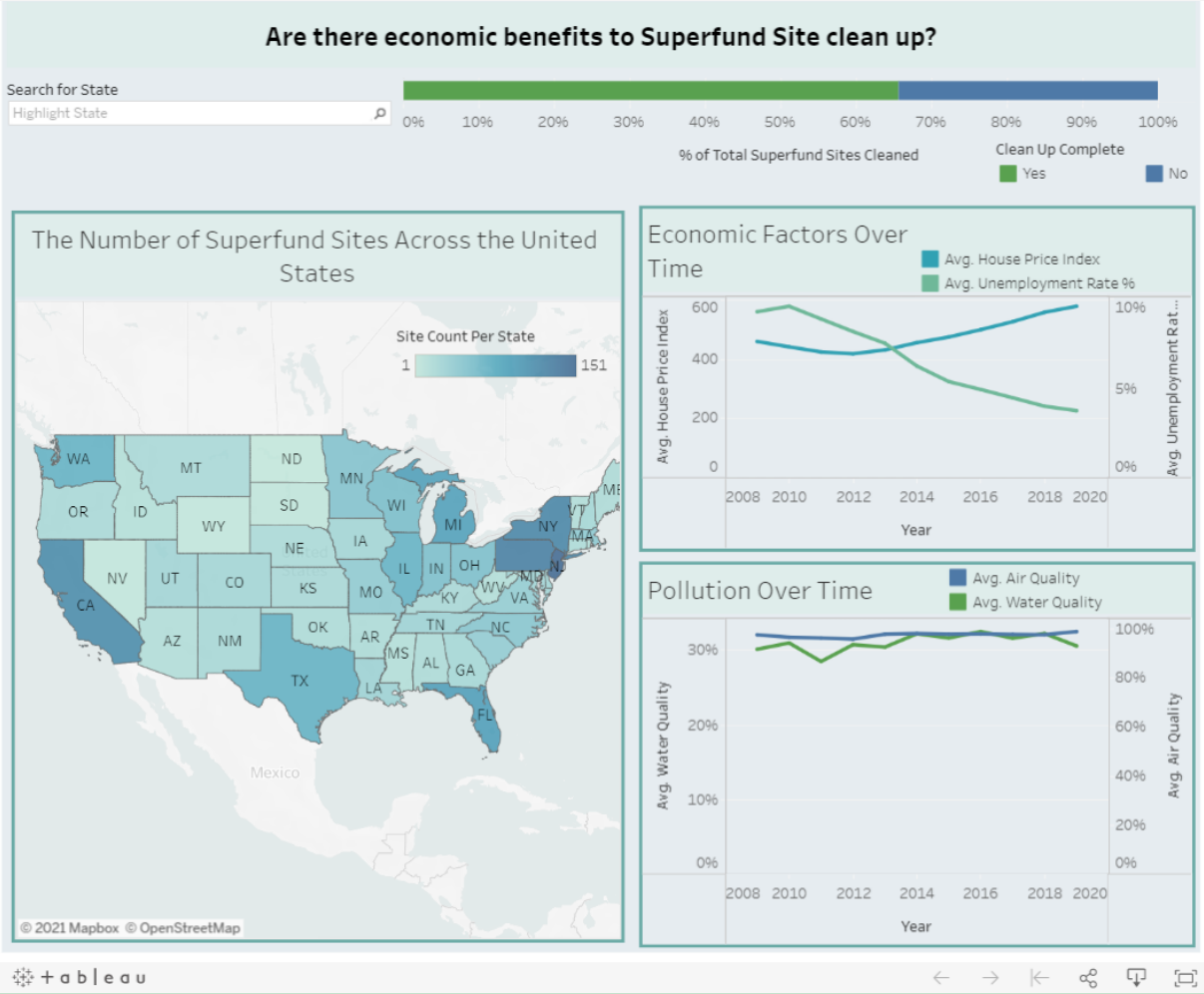
Dep. Variable:	unemployment_lag_1	R-squared:	0.067
Model:	OLS	Adj. R-squared:	0.062

	coef	std err	t	P-value
Intercept	9.4795	1.017	9.322	0.000
water_quality	0.0041	0.002	1.960	0.051
air_quality	-5.5232	1.043	-5.296	0.000

DESCRIPTION OF DASHBOARD

Our project and research question did not lend itself to an end product for a client, so was developed as an exploratory tool instead. We were able to use the dashboard to help us further understand our data as well as the impact and the relationships between our variables of interest.

Using the dashboard, we could review our two case studies of California and New Jersey. We zoomed in to their data and studied the relationships and correlations at the county level. Users could use the dashboard to take a closer look into the Superfund sites, pollution, and economics of their own state and county.



The finalized dashboard was developed in Tableau using the full dataset. It is located at Tableau Public. It includes a heat map that shows the amount of superfund sites in each state. It also has visualizations on the side of the map that display data over a range of ten years from 2009-2019. There are two line graphs with a dual axis: one for economic factors over time (house price index and unemployment rate), the other one for pollution over time (air quality and water quality).

On the upper right hand corner is a bar graph showing the percentage of sites that are cleaned and deemed safe according to EPA standards. In the upper left hand corner is a search bar to zoom in on a particular state. Before interacting with the map, the visualizations show national average data. Once the user clicks on the state, the dashboard swaps views to the second part of the dashboard. This view will show the user state and county level data. As mentioned, this second view shows state level data before interacting with the heat map. Once the user clicks on a county, it updates the graphs.

CONCLUSION

In this study, we explored air and water pollution as well as economic factors to answer two main research questions: 1) How has the 1980 establishment of the Superfund affected pollution?, and 2) What have been the economic impacts?

Pollution has decreased over time.

We found that both water quality and air quality have improved over time from 2009 to 2019.

Cleaned sites have better economic welfare and pollution statistics.

Cleaned superfund sites were correlated with better air and water quality, lower unemployment, and higher house prices.

Air and water quality are not reliable predictors of economic welfare.

Using cleanup to predict economic factors was less reliable, however. When trying to use economic models, we were unable to say that the pollution variables we chose could reliably predict our chosen economic variables.

FUTURE WORK

It is important to understand some of the ways our study was limited.

01 — Economic Variables

First, there are many factors that affect unemployment and house prices. Identifying factors that have macroeconomic effects would allow us to build a more comprehensive model.

02 — Contamination Mediums

Second, in our study, we use air and water quality variables as indicators for Superfund cleanup. However, the contaminants from Superfund sites may not only affect other and different mediums other than air and water, but also multiple mediums simultaneously.There's an opportunity in future studies to understand a little more about how the contaminants affect different mediums and what those are to again, build a more comprehensive model where we could more directly pair site issues with the effective media.

03 — Individual Level Data

Another limitation that we wish we could have more easily corrected is our lack of data at a lower level. For instance, we had zip code data for the sites, of course, but none of the pollution or economic variables that we had having such data may have allowed us to see the populations that are affected by the Superfund sites, and if there are any differences when we drill down to that zip code level.

04 — Time Period

Finally, due to the limitations in data for some variables, we had to limit the data to a 10-year period from 2009 to 2019, which when adding a lag of one year for our economic variables led to a 9-year period. The effects that cleanup has on pollution and therefore the economic factors may take longer to see.

Studies on Superfund site cleanup often only consider benefits from a health perspective, but there may be a greater macroeconomic effect. We encourage future studies to continue to explore site cleanup from this economic perspective. We would recommend that future work focus on building out a more comprehensive model with consideration of our listed limitations.

ACKNOWLEDGEMENTS

Those responsible for concept and coordination:

Reed Thunstrom | Mentor

Taylor Isom | DS4A TA

David Hagmann | DS4A TA

Joycelyn Streator | Data Analyst

Russell Wasem | EPA consult

We appreciate those who have supported this work and helped guide this project.