Creating a Predictive Model for Potential Hazardous Locations

By

Christopher Smith

CSC 498-02 / Spring 2021 Mentored Research:
3/26/2021
Instructor: Dr. Pulimood

ABSTRACT

This paper focuses on the challenges of predicting possible brownfield locations in New Jersey. An interactive mapping approach is made that visualizes the locations of active brownfield sites. Based upon the location of these sites, different methods are proposed to create an accurate predictive model of potential hazardous locations. This model is needed in order to save resources for organizations that acquire property without knowing if the land contains hazardous chemicals or not. Different predictors such as chemicals, proximity to nearby sites, and frequency of nearby sites can be used to predict potential brownfield locations. There are some issues with the approach taken but they theoretically can be solved.

Keywords

Brownfields, Estimation, Predictive Modeling, Clustering, Interactive mapping

1. INTRODUCTION

1.1 Background

To define brownfields, “brownfields are sites that are in need of revitalization and are environmentally compromised sites [Bross and Boyle 2017].” These polluted areas are created by “factories, gasoline stations, dry cleaning establishments, chemical storage companies, and former landfills [Bross and Boyle 2017].” These sites are expensive to clean up, and some sites that are heavily polluted are not listed as a brownfield on the NJDEP’s active sites list. This makes it expensive for agencies that rehabilitate areas that are heavily contaminated. One example of this would be the nonprofit organization named ‘Habitat for Humanity’. This organization full of volunteers is in charge of building and improving homes in order to create affordable housing to strengthen and stabilize surrounding communities. When redeveloping property, this organization doesn’t know which properties have chemicals in the soil that require cleanup. This can lead to expensive cleanup and put the volunteers at risk for coming in contact with certain chemicals when rebuilding the area.

1.2 Solution goals

In order to solve this issue, a predictive model would be used to predict sites that could contain hazardous chemicals. This model would be based upon different predictors such as the chemicals in the soil, proximity to other brownfields, and other factors. The NJDEP has a list of labeled brownfields where some sites are regulated and dispose of waste in nearby areas. These are listed as active sites, where the location of these sites is listed as an address found within a county or municipality. Contamination spreading from these active sites could lead to property that is not suitable for living conditions within the surrounding area. Unfortunately there are many areas like this in the Trenton area. Since cleanup is expensive for cleaning contaminated land, these

nonprofit organizations would benefit from a technique that will provide information before they clean up. A model will be created in order to avoid acquiring land that may be contaminated. This model is intended to include not just the Mercer County area but will be able to be expanded to include more counties in New Jersey. This model would provide a heatmap from yellow to red where areas in yellow could potentially be hazardous locations and red areas are likely or listed brownfield areas. By providing a heat map for this issue, organizations will be able to steer clear of possible hazardous locations that will take up resources to clean up.

2. ISSUES

2.1 Mapping to a Geographic Location

In order to create an accurate representation of the Mercer County Area, a geomap is being implemented that contains the longitude and latitude coordinates of contaminated or potentially contaminated sites. However, when converting the address of a site into a geographic location by implementing an autonomous converter, the address could be converted to the wrong location. This happens because there could be two identical addresses that are in different states, or there could be similar addresses within the same state.

This problem can be resolved by adding the zip code of the area as part of the dataframe. It is unlikely that two addresses will be found within the same zip code.

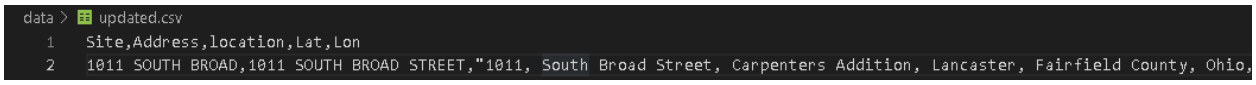

Figure 2.1 Address in New Jersey is mapped to a similar address in Ohio

2.2 Availability of information

It is difficult to find updated information for brownfields that contains chemical information, zip code, and address. By using the DEP DataMiner [DEP 2016] it is possible to find a list of active sites in a specific county. However the site name and address are the only characteristics that would be beneficial for the predictive model. This data can be extracted into a csv file using a web crawler but more information is needed in order to create an accurate model.

3. DESIGN

3.1 Collecting the data

The first step to creating this model is to collect information about the addresses of the known sites. This information can be found using the DEP DataMiner [DEP 2016]. In order to gather this information a web scraper was implemented in python that would navigate to the NJDEP dataminer active sites website. Then it searches for active sites that are located in Mercer County. After the page is loaded, the address fields are pulled from the website into a csv file that is labeled addresses. This web scraper was implemented in python using the selenium package.

3.2 Converting address to location coordinates

Once the data is collected the addresses need to be converted to the coordinates of the location. This can be done by using a geocoder to convert physical addresses to geographic locations [Abdishakur 2019]. This is necessary in order to visualize these locations on a map. By using the

geopy package in python these locations are converted into latitude and longitude coordinates. This example uses Nominatim but there are other geocoders available. In figure 3.2 location, latitude, and longitude are saved as columns for the new dataframe that we created that will display the coordinates in a new csv file named updated.csv.

coordinates.py > df

```python
from geopy.geocoders import Nominatim
from geopy.extra.rate_limiter import RateLimiter

import pandas as pd
#Read in addresses csv file
df = pd.read_csv("NJPDES original.csv")
#choose geocoder
geolocator = Nominatim(user_agent="myGeocoder")
geocode = RateLimiter(geolocator.geocode, min_delay_seconds=1)
#Save as location column in dataframe
df['location'] = df['FACILITY'].apply(geocode)
#Save as latitude column in dataframe
df['Lat'] = df['location'].apply(lambda x: x.latitude if x else None)
#Save as longitude column in dataframe
df['Lon'] = df['location'].apply(lambda x: x.longitude if x else None)
#df=df.dropna(subset=['location'])
#Save as a csv file
df.to_csv('updated.csv', index=False)


```

Figure 3.2 Example code of converting brownfield addresses

3.3 Visualizing the Data

In order to visualize this data with the coordinates that are now saved in the updated csv file a map will be created with markings of all the active brownfield sites. In this example leaflet was used with JavaScript to mark a sample of the locations [Agafonkin 2021]. This package uses OpenStreetMap to create an interactive map that the user can zoom in and out of. In the figure below an example of the locations for active brownfield sites are plotted. However, some addresses were mapped to the wrong location which can be seen since there are mappings for locations in other states.

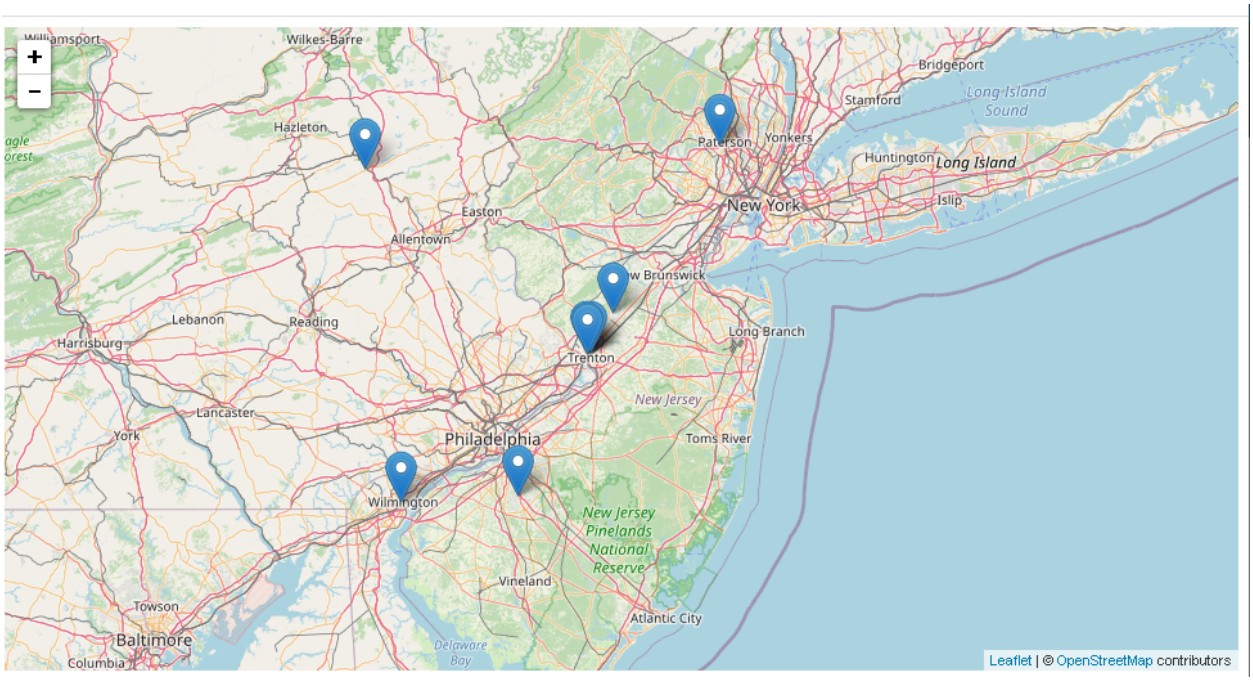

Figure 3.3 Interactive map example for active sites

3.4 Implementing a Predictive Model

To create an accurate predictive model different predictors can be used in order to provide more accurate estimates. One approach to predicting locations that could be potentially hazardous would be using the cluster approach [Charfaoui 2019]. This approach creates areas that contain a significant amount of points together. This would make it possible to define areas that contain a significant number of active brownfield sites and label surrounding areas as potential hazardous locations. This can be implemented through hierarchical clustering since some clusters may contain more brownfields than others. However this model takes into account the distance between different locations so this would have to be defined beforehand in the csv file.

4. RESULTS AND ANALYSIS

4.1 Collecting the Results

The number and grouping of clusters were able to be collected through the use of a DBSCAN(Density Based) model. This model takes into account two parameters, which are “the physical distance from each point and a minimum cluster size [Boeing 2014].” The neighborhood is the space that is defined around a data point, in this case our points are locations and the neighborhoods are the maximum space between locations. In the model the radius of the neighborhoods is set to .2 km and the minimum number of data points is set to one for cautionary zones and three for more likely zones to be contaminated. The yellow zones will be set for areas that could be contaminated while the red zones are areas that are likely to be contaminated.

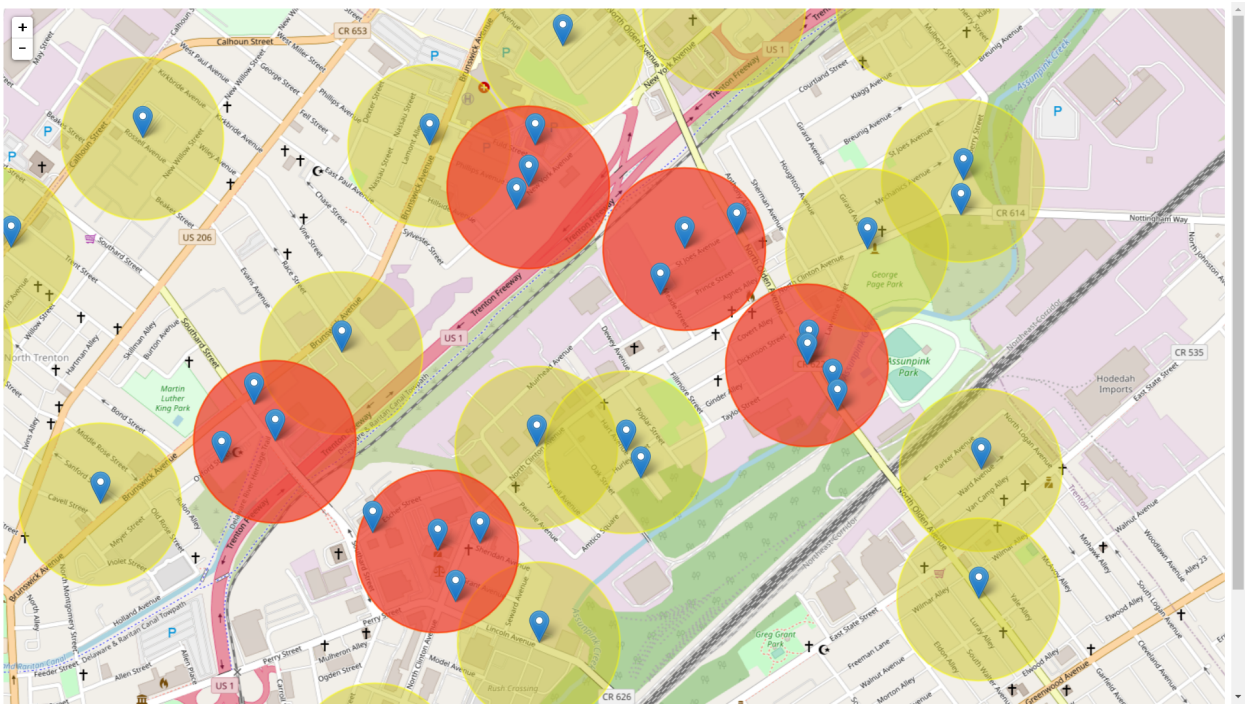

Figure 4.1 Clustering Results with Possible Contaminated and Likely Contaminated Zones

4.2 Analyzing the Results

As shown in Figure 4.1, the red zones which contain areas that are likely to be contaminated contained three locations at minimum. Overall, there were ten clusters for likely contaminated zones and there were sixty clusters found for likely contaminated zones for the Trenton area. There were many more clusters that covered a larger area than I originally predicted. All these clusters were just for the Trenton area as well, displaying that there are many areas that buyers should be weary of. Before buying property located in these possibly contaminated areas, the landowners should test the soil and estimate the cost of cleaning the land. Also, property buyers could consider buying land that is not found in these zones in order to save estimated costs on decontaminating their land. Keep in mind that these are estimations, and that the land found on these zones could be contaminated or not. Also, land that is not found within these zones may be contaminated upon further inspection of the property and its soil.

5. CONCLUSION

Overall, the clustering algorithm generated zones, but the user should be careful about inferring conclusions on the data found. The results found use a prototype so further testing should be completed before using this data for acquiring land. Future research containing the chemicals in the soil could be provided to further improve the accuracy of the clustering algorithm. Also, the data collected could be extended for other locations beside Trenton in order to cover more areas within New Jersey. The method of collecting data can be expanded since it can be changed to scrape data from other municipalities besides Trenton. The locations marked on the map using OpenStreetMap can be expanded to include other areas besides Trenton as well.

6. BIBLIOGRAPHY

NJDEP. 2020. *Site Remediation Program* (CPATH). https://www.state.nj.us/dep/srp/kcsnj/#:~:text=Active%20Sites%20are%20those%20sites,pending%20and%2For%20closed%20cases. Accessed 4/23/2021.

Bross, P. E., & Boyle, S. B. 2017. *The Greening of New Jersey's "Brownfields" - As Viewed by the Department of Environmental Protection, 9*(3), 1-27. DOI=https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1503&context=elr. Accessed 4/23/2021.

Abdishakur. 2019. *Geocode with Python* (CPATH). https://towardsdatascience.com/geocode-with-python-161ec1e62b89. Accessed 4/23/2021.

Agafonkin, Vladimir. 2021. *Leaflet Quick Start Guide* (CPATH). https://leafletjs.com/examples/quick-start/. Accessed 4/23/2021.

Charfaoui, Younes. 2019. *Working with Geospatial Data in Machine Learning* (CPATH). https://heartbeat.fritz.ai/working-with-geospatial-data-in-machine-learning-ad4097c7228d. Accessed 4/23/2021.

Boeing, Geoff. 2014. *Clustering to Reduce Spatial Data Set Size* (CPATH). https://geoffboeing.com/2014/08/clustering-to-reduce-spatial-data-set-size/ Accessed 4/23/2021.