Identifying Vacant and Uninhabitable Properties in Philadelphia using Publicly Available Data

By Akshay Srivastava, Prianka Ball, Reina Carissa

DS 670 - Capstone: Big Data & Business Analytics

Abstract

This study presents an approach to predicting vacant lots at the lowest granular level possible in the City of Philadelphia using publicly available datasets. We use a variety of data that have been historically connected with vacant lot issues such as 311 calls, crimes, property tax delinquency, and property violation. Our research incorporates block group level data collected by the American Community Survey and data collected by the City of Philadelphia. We use geospatial datasets as well as regular datasets and have achieved 97% precision and 96% recall for predicting vacant lots.

I. Introduction

The case in urban vacancy

Urban vacancy is a contested issue in many cities across the U.S and globally (Nassauer & Raskin, 2014). Population migration has left vast swaths of abandoned houses and factories as manufacturing jobs were either automated or operations moved to cheaper labor markets (Esposito, 2020). Gentrification is transforming communities and encouraging people to move to America's larger cities (Favre, 2019), leaving many others to experience decentralization, deindustrialization, and population decline in the past decades (Pearsall et al. 2013). Over time, these vacant properties decreased in value due to neglect and now represent the empty lots and buildings we see today.

Vacancy produces negative economic impacts and is associated with many concerns among communities (Accordino & Johnson, 2000). The longer a property sits vacant, the greater the risk that it will deteriorate. Vacant properties are strongly associated with crime, violence, vandalism, illegal dumping, and arson (Mallach, 2018). These resulted in a decrease in standards of living, public health, and other negative indicators of community well-being.

Furthermore, urban vacancy can significantly reduce the value of the surrounding occupied properties (Mallach, 2018). According to a study conducted by the Temple University Center for Public Policy (2003), a vacant building on a block can reduce the value of nearby properties by 20 percent or more, which greatly impacts the costs for local government. They not only reduce local property tax revenues due to their low value, but also cost cities millions of dollars for policing, inspecting, cleaning vacant lots, and demolishing derelict buildings (Mallach, 2018).

The longer a property remains vacant, the greater its impact on surrounding property values and the larger the radius of this effect. If an area has too few vacant units relative to demand, prices may rise unreasonably because of the shortage of supply; if there are too many, oversupply may push prices and rents down to the point where homeowners find themselves underwater, and landlords may not make enough money to cover their costs (Mallach 2018). The Office of Policy Development and Research (PD&R; 2014) research also suggests that a vacant property increases the rate of violent crime within 250 feet of the property to be 15 percent higher than the rate in the area between 250 and 353 feet from the property. When vacancies rise above approximately 20 percent of an area's total properties, the number of vacant buildings and lots may continue to grow indefinitely (Mallach, 2018).

Urban vacancy in Philadelphia

Vacancies are still at epidemic levels in many older cities, particularly in the nation's legacy cities. Like many cities in the Northeast, Philadelphia in the state of Pennsylvania was home to a growing manufacturing economy that spurred population growth, which in turn sparked private investment in the housing market and municipal services (Schilling & Hodgson, 2013), thanks to its manufacturing industries that employed nearly 50 percent of their workforce.

By the mid-20th century, however, Philadelphia lost over half of its industrial sector, approximately 160,000 manufacturing jobs (Schilling & Hodgson, 2013), and experienced a significant population loss, like most older industrial cities in the 20th century. Philadelphia's population declined steadily in the second half of the 20th century, dropping from a high of just over 2 million in 1950 to a low close to 1.5 million in 2000 (Chowdhury, 2017). Population trends started to shift as Philadelphia started to attract new resident markets, especially immigrants and empty nesters.

Philadelphia is currently the 6th largest city in the United States, based on a population of 1.6 million according to the World Population. With close to 40,000 estimated vacant lots, this cost the city $20 million in maintenance costs and $2 million in uncollected property taxes every year (Econsult Solutions, 2010). A study found that houses within 150 feet of a vacant or abandoned property in Philadelphia experienced a net loss of $7,627 in value (Bass et al. 2005). In the case of the city of Philadelphia, the management, clean-up, disposition, and redevelopment of vacant land is scattered across 15 separate city departments, each of which has different policies and objectives (Econsult Solutions, 2010).

With three-quarters of vacant lots in Philadelphia being privately owned, unfortunately not all owners take the necessary steps to protect and care for their property. Many of these lots have become sites for illegal dumping for soiled mattresses, abandoned cars, and household trash (Loesch, 2020). Overcoming the vacant property problem in Philadelphia will eventually require a similar focus on gathering and using information about both public and privately owned properties.

A. Srivastava
P. Ball
R. Carissa

Addressing vacant lands in Philadelphia

Cities have adopted a broad range of policies to address the vacant land problem with varying degrees of success since the 1950s (Pearsall et al., 2013). Fixing up abandoned land has shown to increase property values of the surrounding neighborhood by 17 percent, according to a report from the Wharton School of Business of the University of Pennsylvania (Wachter & Gillen, 2005). Another study has shown that remediating abandoned buildings and lots reduces gun violence by 39 percent (Branas et al., 2016). The study estimates that each dollar spent repairing vacant land yields a direct $26 rate of return to taxpayers. The study also concluded there were indirect savings of $333 per dollar spent through ripple effects such as reduced violence.

Philadelphia has a long history of experimenting with programs and policies to address its significant vacant property problems. Philadelphia's innovative solution to this demoralizing problem has been initiated both within and outside the city government (Schilling & Hodgson, 2013), but the problem remains the same–while national source exists that projects the number and location of vacant lots in a particular location, the data is largely outdated (Mallach, 2018). Many types of vacant property are not measured except when people walk or drive block by block to count them.

As researchers experiment with development ideas of formerly vacant and abandoned properties, they will need to evaluate strategies and determine the most cost-effective and sustainable effort (The Office of Policy Development and Research, 2014). More research will help decision-makers become better equipped to turn problem properties into assets that will stabilize and revitalize neighborhoods and improve residents' quality of life, but unsurprisingly, this cannot be done unless accurate data is used in the decision-making process.

More research will be needed to empower policymakers, investors, and citizens to make evidence-based decisions on difficult choices, such as when to rehabilitate and when to demolish, whether to have a judicial or administrative foreclosure process, whether to convert a brownfield to an affordable housing development or a green space, or whether a particular area should pursue smart growth or smart decline.

Research objectives

This study aims to assist decision-makers using machine learning (ML) models to accurately identify lots that are in danger of becoming vacant by incorporating publicly available data on City Government websites and the American Community Survey (ACS). We will be exploring data that have been historically connected to vacant lots such as crime rate, property assessments, 311 calls, property tax delinquency, property code violations. We have treated this problem as a classification problem. Vacant lot data available on property assessment dataset was used as the labeled dataset to train the model. Through geospatial analysis and the integration of data from multiple sources, we manage to detect vacant lots early in the process that can improve quality of life and revitalize disinvested urban neighborhoods in Philadelphia.

A. Srivastava
P. Ball
R. Carissa

II. Datasets

Data collection

Various types of data are accumulated and compiled in a list of data sources that were used to predict and access vacant lots. We have used different types of packages. For our project, GeoPandas and CensusData packages played a key role in our analysis. We used GeoPandas to map, join and analyze geospatial data. CensusData package was used to easily pull ACS data. Most of the datasets we have used were collected between 2015 and 2021.

This project collected data from the publicly available resources, as listed below:

1. **Philadelphia shapefiles**: shapefile is a data storage format for storing the location, shape, and attributes of geo. Shapefiles assist us in creating maps and joining with other types of datasets.
    - Census Block Groups: Census block group is a geographic unit used by the US Census Bureau. It is the smallest geographical unit for which the bureau publishes sample data. Typically Block Groups have a population of 600 to 3,000 people. Data collected from ACS had data at the census block group level, so this shapefile was important to analyze this dataset. (Source: https://www.opendataphilly.org/dataset/census-block-groups)
    - ZIP Code: ZIP Code is a postal code used by the United States Postal Service (USPS). The basic format normally consists of 5 digits. Some of the datasets collected from the city government have ZIP code level data that can be plotted as a map. (Source: https://www.opendataphilly.org/dataset/zip-codes)
2. **American Community Survey (ACS)** is conducted by the Census Bureau. Unlike the Decennial Census which is conducted every 10 years, the ACS is conducted more frequently. For our study purposes, we used the 5-year ACS data where the data has been collected over 5 years between 2015 to 2019. Data in ACS is at the block group level. CensusData package was used to pull the data.
    - Occupancy Status (Table Code: B25002): Data on whether the property was vacant or occupied.
    - Vacancy Status (Table Code: B25004): Vacant properties could be further broken down according to their housing market classification. For our purposes, we will be focusing on the “other” vacancy status. Vacant status is classified as other when it does not fall in any of the year-round categories.
    - Population: Population data was also pulled from ACS to understand population density in each block group.
3. **Philadelphia city**: The Open Data Program of the City of Philadelphia helps

departments share data from the city government with the public on OpenDataPhilly.org. We used the following data from Philadelphia City.

- *Property Assessment* data is collected by the Philadelphia Properties and Assessment History. It includes property characteristics and assessment information from the Office of Property Assessment. This dataset includes data of properties that are already known as vacant lands by the city government. (Source: https://metadata.phila.gov/#home/datasetdetails/5543865f20583086178c4ee5/)

- *Property Tax Delinquency* data shows Philadelphia properties with reported tax delinquencies, including those that are in payment agreements. An account is considered delinquent when Real Estate Tax is still unpaid on January 1 the following year the tax was due. The data collected is from 1972 to 2018. Vacant lots tend to have unpaid taxes and this is a good indicator that properties might be vacant soon. (Source: https://metadata.phila.gov/#home/datasetdetails/57d9643afab162fe2708224e/representat)

- *Property Code Violations* data contains violations issued by the Department of License and Inspection. Data contains where the violation occurred and the reason for the violation. Some of the violations are vacant lot-related. (Source: https://www.opendataphilly.org/dataset/licenses-and-inspections2violations)

- *Crime*. Due to the extent of the impact, crime can work as an indicator of vacant lot location prediction. The data is collected from the Philadelphia Police Department. (Source: https://metadata.phila.gov/#home/datasetdetails/554386892058308617 8c4f8e/representat)

- *311 Data* represents all service and information requests since December 8th, 2014 submitted to Philly311 via the 311 mobile application, calls, walk-ins, emails, the 311 website, or social media. We incorporated this data set as vacant lots tend to get more 311 calls to reports about things like illegal dumping, maintenance services, or graffiti removal. (Source: https://metadata.phila.gov/#home/datasetdetails/5543864d20583086178c4e98/representa)

Exploratory data analysis (EDA)

The libraries we used include GeoPandas for geospatial data, CensusData for census data, Sklearn for ML model and feature engineering, Seaborn and Matplotlib for statistical visualization, Shapely, Pandas, and NumPy.

Using the ACS data, we narrowed down the census data to the last 5 years specifically in

A. Srivastava
P. Ball
R. Carissa

Pennsylvania with a focus on Philadelphia county. Based on the vacant lot occupancy data collected by the ACS, we have noticed that the median percentage of vacant places in each block group is 12 percent, calculated using total vacant places divided by the total places for each block group. Our study has shown that there are different types of vacant places, which range from ‘vacation spots’ to ‘vacant due to switching tenants,’ but the majority of the actual vacant lot according to our definition would fall under the “other” vacant type. The percentage of this type of vacant place is around 6.7 percent, calculated using the “other” type number divided by the number of total places for each block group.

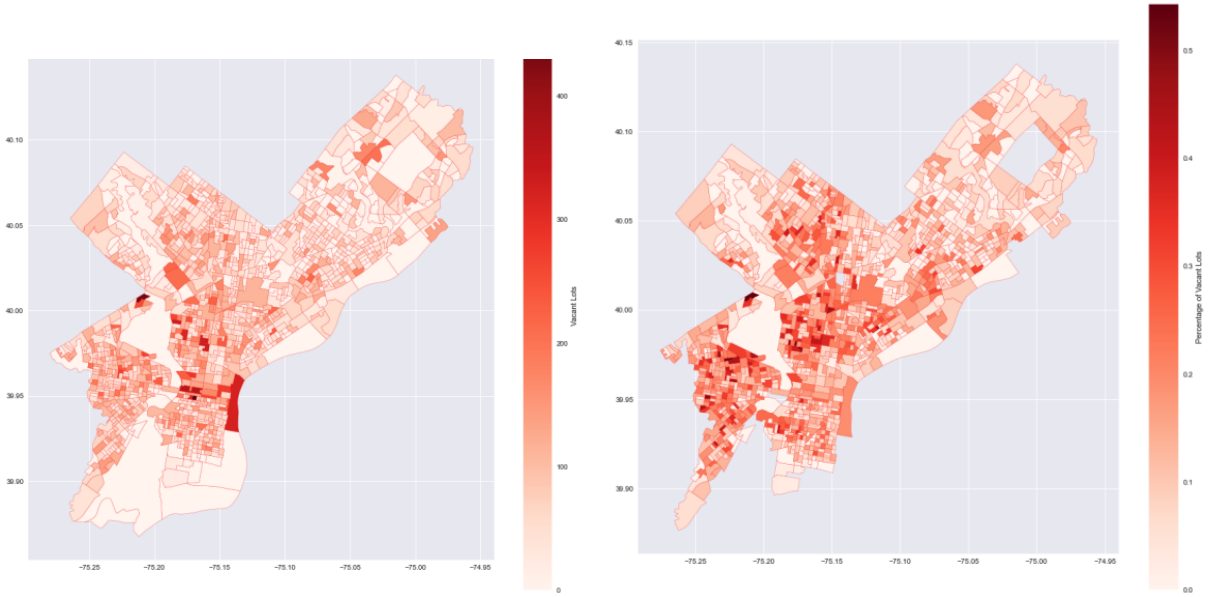


Figure 1 & 2. The number of vacant places in each block and the percentage of vacant places in each block group.

The map shown for the percentage of the vacant places is different compared to the actual number of vacant places because the percentage is relative to the total number of places in the block group. As an example, the busiest area in Philadelphia may have fewer vacant lots due to the rapid development, compared to the more suburban areas, which are not necessarily “vacant” due to abandonment. The percentage here takes into account the density of livable places in the block group.

One of the most complicated datasets to understand was the Property Tax Delinquency dataset. When property owners do not pay the full amount of state or local property taxes assessed against the value of the property, their properties are considered tax-delinquent (“Foreclosure and Disposition of Tax-Delinquent Properties”). We have included a lot of the variables in our final model from this dataset so it was very important for us to understand the context. Some of the important variables from this dataset are described in the following section.

Property tax delinquency data is divided into two categories, actionable and non-actionable. If the city is actively working to collect accounts, the properties are considered to be ‘actionable.’

A. Srivastava
P. Ball
R. Carissa

Properties become ‘non-actionable’ when the city is unable to collect accounts or barred from doing so. Accounts that are in a payment agreement, bankrupt, or overdue but not yet delinquent are considered non-actionable (“Real Estate Tax”). A payment agreement is one of the methods for a city to collect debts.

Sheriff’s sale and sequestration fall under the actionable category. A sheriff's sale is a public auction where mortgage lenders, banks, tax collectors, and other litigants can collect money lost on the property (Chen, 2021), while sequestration is the act of taking of someone's property, voluntarily or involuntarily, by court officers or into the possession of a third party, awaiting the outcome of a trial in which ownership of that property is at issue (“Sequestration,” 2017).

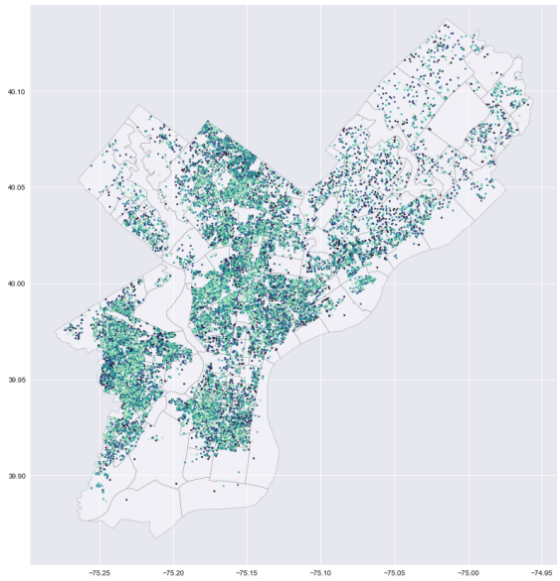


Figure 3. Property tax delinquency of principal value more than $2K on the neighborhood. The lighter the color, the higher the principal value.

Based on the data, most of the delinquent properties are residential and are owed for 1-6 years on average, some are also owed at 25 years. When looking at residential and commercial properties, we noticed that as much as 89 percent of the properties are residential and delinquent. If the delinquent property is a rental property, the city can take over the rent collection and apply those rental payments to the delinquent Real Estate Tax bill.

Houses and vacant lands have the highest principal due, which is the amount of principal that has come due on a loan minus the amount that has been paid. Most of the principal due is owed for 1 to 4 years. A significant portion of the principal due was also owed for 25 years. When looking at the median principal due, a lot of taxes are owed for 18, 23, 22, and 27 years. Since the principal value is skewed, we use the median at around $2,000 as a measure to analyze the dataset.

Another dataset we take into account is property code violations. These codes are important in

order to enhance and preserve the quality of neighborhoods and working environments. Typical violations of the codes include junk, trash, or debris on private properties, inoperable or abandoned vehicles on private properties, graffiti on private properties, weeds & overgrown vegetation, and substandard and/or unsafe structures (“Code Compliance”).

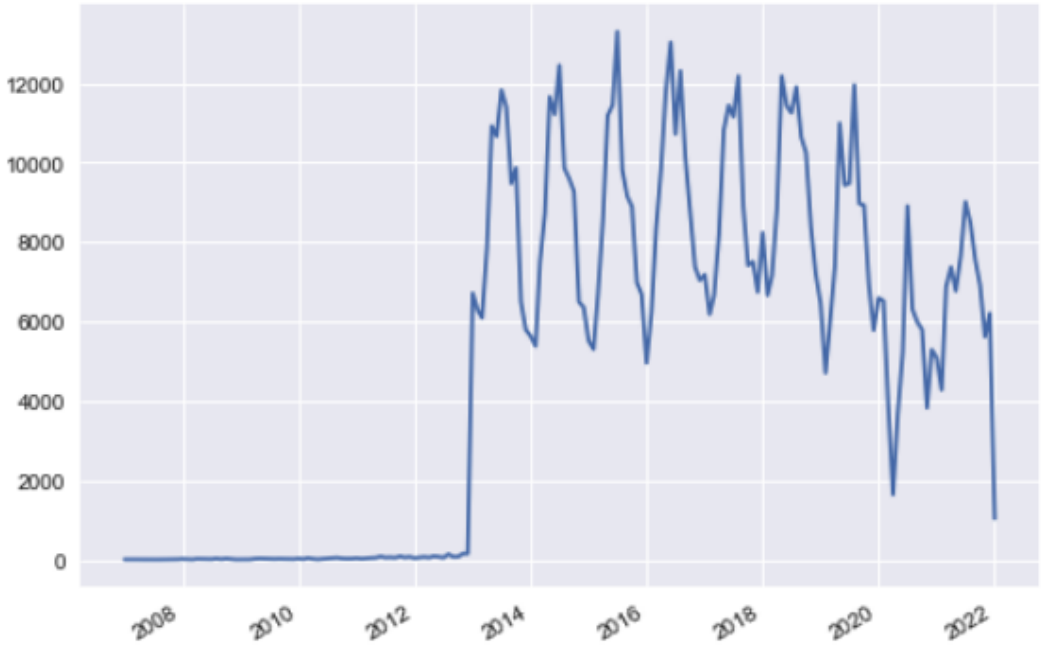


Figure 4. Number of reported case violations year 2015-2021

There are close to 2,000 types of code violations. From the above graph, we observed some seasonality in code violations, where it seemed to peak towards the end of the year or wintertime, where there may be some added complications due to weather conditions.

Abandoned buildings have become a symbol that a neighborhood has deteriorated (Branas et al., 2012). Aside from unmaintained lots that result in code violations, abandoned buildings may also contribute to issues related to crime rates. Vacant lots offer an ideal hideout location for criminal and other illegal activity.

A. Srivastava
P. Ball
R. Carissa

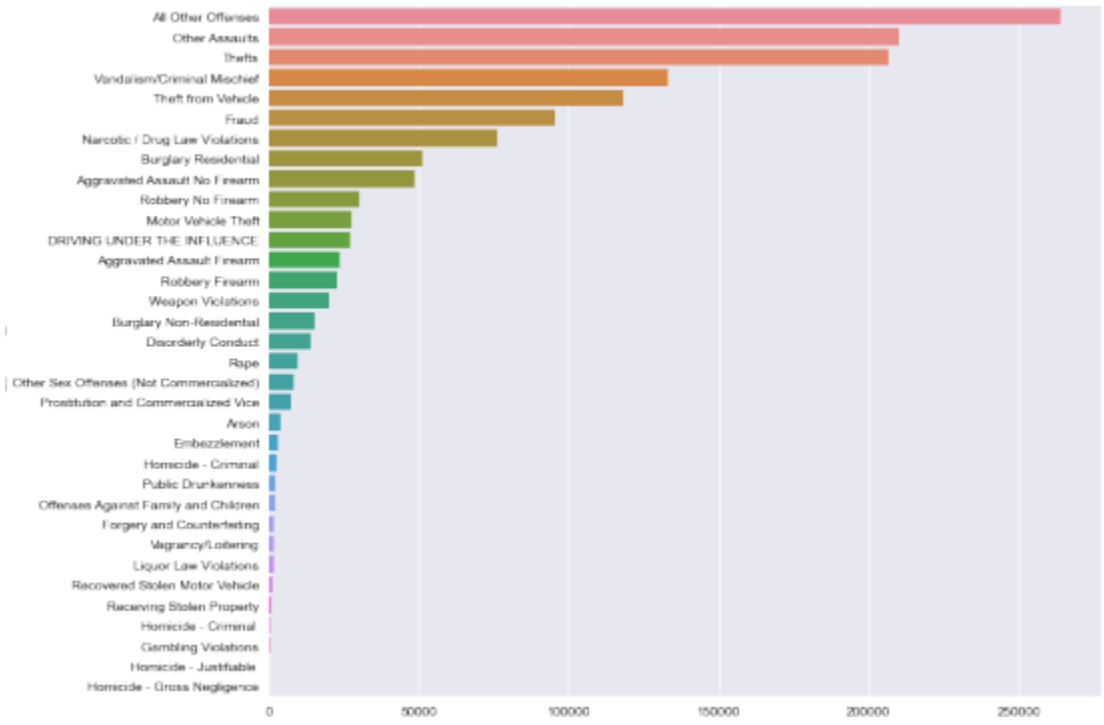


Figure 5. Most of the criminal cases falls under “all other offenses.”

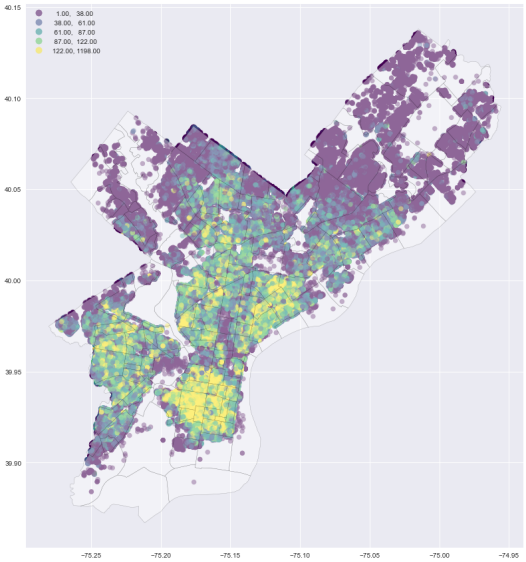
The majority of reported crime resulted from “all other offenses” that includes all violations of state and local laws not specifically defined as a Part I or II offense, except traffic violations. They are still considered emergencies but do not fall under the category determined in the dataset. After grouping the number of crimes that happened within 50-meter of each `parcel_number`, we have noticed crime data to be skewed.

The 311 calls are easy-to-remember telephone number that connects customers with highly-trained customer service representatives ready to help the community with non-emergency affairs. Although a large portion of the latitude and longitude data were missing from this dataset, the summary of this data can be visualized below.

A. Srivastava
P. Ball
R. Carissa

Figure 6. Majority of the 311 calls are attributed to “information requests.”

Similar to the crime data, after grouping the number of 311 calls that happened within 50-meter of each `parcel_number`, we noticed skewness in the dataset. After plotting 311 calls and crimes that occurred within 50m of each parcel number, we have noticed that the areas that tend to have a high number of 311 calls and crimes also tend to have higher concentrations of vacant lots. Most of the datasets that we collected were very skewed.



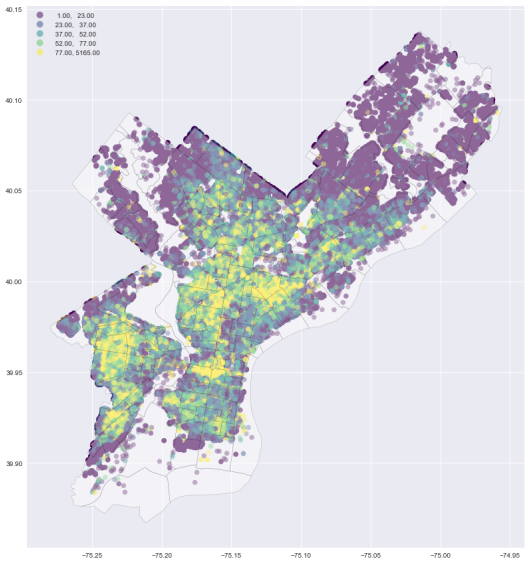


Figure 7 & 8. 311 Calls within 50-meter of each parcel number and crimes within 50-meter of each parcel number.

A. Srivastava
P. Ball
R. Carissa

III. Methodology

We used a binary classification approach to predicting vacant lots. The `category_code_description` variable from the Property Assessment dataset was used to create labeled column Y for ML models. But before we could start working on the model, we had to join the different datasets, clean the data, and include more variables that would add additional information for the model to run smoothly.

Data Preprocessing and Feature Extraction

To connect the different datasets, we primarily used `parcel_number` or `opa_number`, which is a nine-digit parcel identifier/account number created by the Board of Revision of Taxes Staff to identify a specific property. This is used specifically to connect the *Property Assessment*, *Property Tax Delinquency,* and *Property Code Violations* data. Since the *Property Assessment* dataset has the largest number of distinct `parcel_number`, we used this as the parent on which the rest of the datasets are joined. To combine the ACS dataset, we use the block groups and spatially joined with the rest of the datasets using the latitude and longitude of the properties in a certain block group.

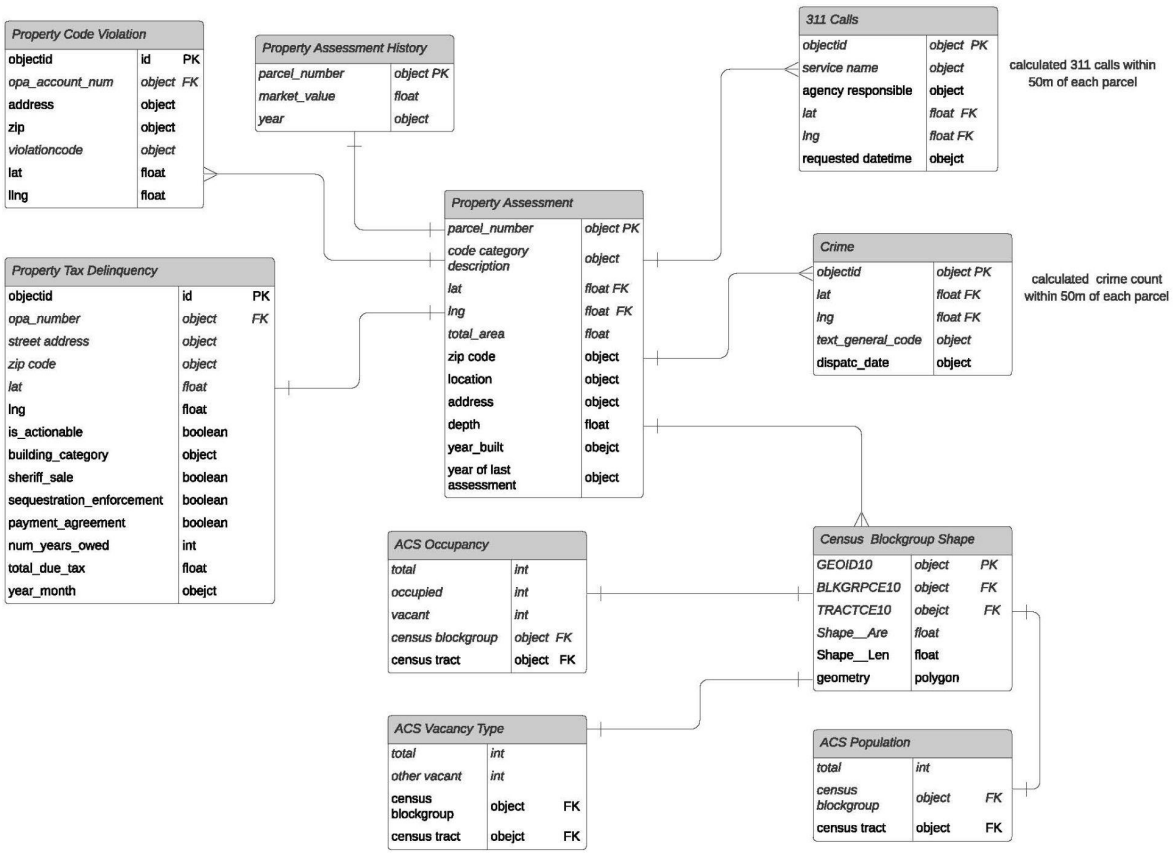


Figure 9. Data connection diagram that shows the most important variables extracted from different

datasets to be joined together. Datasets contained more variables than the ones shown here.

`Parcel_number` is a piece of location information used to identify the property for tax, title, deed, and property line reasons. Since the crime and the 311 data only contains location data and not `parcel_number`, they are joined to the parent table by tallying the number of crimes and 311 calls made within 50 meters of each `parcel_number` after we took into account the average radius between houses in Philadelphia. A similar method to join 311 calls and crimes data with `parcel_number` was used in another paper (Reyes et al., 2016), where 50-meter was the lowest radius they used. For our purposes, we decided to only use a 50-meter radius as Philadelphia is more densely populated than Cincinnati.

We used three different datasets from the ACS that were joined together and later were used to create additional columns. From the original dataset, we included total housing, “other” vacant, total vacant, and total population. We decided to leave out the different types of vacancies except for the “other” vacancy as they were not connected to the issue we are working on. In addition to these variables, we have also created percentage vacant lots of total places in each block group, percentage “other” vacant lots of total places in each block group, and population density per km^2 in each block group. The population density variable was created after joining ACS datasets with the block group shapefile. The block group shapefile gave us access to the area of each block group which was used to create population density. Other than area, the shapefile also gave us access to block group-specific information such as GEOID, block group number, tract number, and perimeter.

The property code violation dataset contained important information about the type of violation that occurred in each property. Each parcel number in the dataset had multiple code violations so we had to find a way to extract the information that would be most helpful for the goal of the project. We created new columns such as the `num_vacant_code` which calculates the number of times an `opa_number` had a vacant lot related violation, a column named `casenumber_diff` which tallies the number of cases in each `opa_number`, a column named `casecreateddate_year_diff` which tallies the different years that appeared for each `opa_number`, `violationcode_diff` which calculated the number of different violation codes that appears for each `opa_number`.

As the latitude and longitude data was very important to join the crime dataset with the rest of the data, we had to drop rows that did not have any location data. As our project only focused on Philadelphia, we removed data that fell outside of Philadelphia. Once the data was cleaned, we segmented the data based on time. Data were segmented into crimes that occurred in the last 6 months (Jul 2021 to Dec 2021), crime in the last 3.5 years (Jul 2018 to Dec 2021), and all crimes (2015-2021). After segmenting the data, they were joined with the property assessment data. The property assessment dataset helped us to understand the location of each property and calculate crimes through three different time periods that occurred within 50m around the property.

A similar approach was taken to calculate 311 calls within 50m of each parcel number. As the

311 calls were collected till 2020, we were limited to only using the last 3.5 years (Jul 2018 to Dec 2020) and the whole period (2015-2020). From our data explorations, we have noticed that a big percentage of the 311 calls were for “Information Requests”‘. As this was not related to vacant lots, we decided to remove it. After its removal, missing latitude and longitude data decreased significantly. Similar to the crime dataset, as latitude and longitude data was very important to join the data with the rest of the datasets, we had to leave out rows with unknown location data.

Current Property Assessment and Historical Property Assessment data were two different datasets. Current property assessment data included very detailed data about each parcel number and the latest information of each parcel number. The historical property assessment data contained the market value of each parcel number in different years. For our analysis, we only included the market value of properties from 2015 to 2021. On the property assessment dataset, we excluded parcel numbers with unknown latitude and longitude as this data was very important for our analysis.

Once all data were combined, some null values were replaced with zero and some with median depending on their feature definition. Number of years owed tax, total tax due, number of property violations, number of different property violations, number of vacant lot related property violations were replaced with 0. Other columns such as the market value of property and depth were replaced with the median.

For categorical columns, null values were not replaced with any values as we were expecting to use models and use feature transformation methods like one-hot encoding that treated null values as a separate category.

Feature Selection

As we were working with multiple datasets that were very big, feature selection was a very important step. Features were selected in two stages. During the first stage, features were selected while going through each dataset separately. Once all datasets were joined, we again went through the process of selecting certain features. During the first step of feature selection, we carefully went through each data description to understand their potential effect on vacant lots that we concluded from our preliminary literature review. If the features had a big percentage of the data that was missing, we decided to exclude them from the dataset. The feature selection process during the second stage was different. During the second stage, we excluded features that had vacant lot data outside of the labeled dataset. We also excluded features that had a high correlation of more than 0.9. The decision to exclude which feature of excluding depended on the feature description. We found a high correlation between crimes that occurred within 3.5 years and the whole period. We decided to exclude crime data from 3.5 years as the whole period data looked more valuable. The same happened with 311 data where there was a high correlation between 311 calls in the last 3.5 years and 311 calls within the whole period. We decided to exclude 311 calls from the last 3.5 years. We also found high correlations between the property market values in different years. We excluded most of the

years except property market values of 2015 and 2021 as the date range was further apart. The correlation between 2015 and 2021 market value was also lower. Principal tax due was highly correlated to total tax due, we decided to exclude principal tax due as it was a lower value. Property code violation in different years was also highly correlated to the number of property code violations. We decided to exclude property code violations in different years and only include the number of property code violations as the later feature seemed to be a more important variable. Other features like bankruptcy were dropped as all variables were the same. Some features like total assessment were also dropped as they were based on other features that were already included in the dataset.

Model Building

For our models, we have used the pipeline method as it lets us use a list of transformers and then a final estimator. Pipelines are set with fit/transform/predict functionality so that we can fit the whole pipeline to the training data and transform to the test data without having to do it individually. Depending on the model that we were going to use, we have used scaler, power transformation, quantile transformation, median imputation on numerical columns. For categorical columns, we used one-hot encoding.

We have tested with different models such as a Logistic Regression, Adaboost with Logistic Regression, Adaboost with Decision Tree, Support Vector classifier(SVC), Random Forest Classifier, XGBoost with Decision Tree. Some of the models were also tested with different transformations. Logistic Regression model was tested with quantile transformation, power transformation, and standard scaler. SVC was also tested with power transformation.

IV. Results

Model Results

As we were working with unbalanced data, we looked closely at the precision and recall scores of vacant lots. We received the best result while using XGBoost which had a 0.97 precision and 0.96 recall score. Outside of XGBoost, we have noticed that other tree-based ensemble models such as Adaboost and Random Forest also tend to perform well. For other models like Logistics Regression and SVC, they only performed well when we used power transformation on numerical columns.

	Vacant Lots		
	precision	**recall**	**f1**
Logistic Regression (Power Transformation)	0.90	0.84	0.87
SVC	0.83	0.97	0.89
Adaboost (Logistic Regression)	0.86	0.45	0.59
Adaboost (Decision Tree)	0.95	0.88	0.91
Random Forest	0.98	0.89	0.93
XGBoost	0.97	0.96	0.96

Figure 10. Fit assessment results from different models.

As we were working with very skewed datasets, it was necessary for the data to have some kind of transformation to make it look more Gaussian, especially while working with models like Logistic Regression and SVC. For SVC to perform well we also needed to use class weight to be balanced as we were working with unbalanced data. For tree-based models, it was not necessary for the data to look more Gaussian.

XGBoost was the top-performing model. Other than a very good recall and precision score, the model also had a high AUC and f1 scores. It had an AUC of 0.992 and an f1 score of 0.964. The model line was also way above the non-skill threshold. Year built was the most important feature that improved the accuracy of the score. Feature importance can be calculated using gain, weight, and cover methods. Their results from our model were very different from each other. 'Weight' is the number of times a feature appeared on the tree, 'cover' is the number of times a feature was used to split the data across all trees and 'gain' shows the relative importance of each feature and shows how much each feature has improved the accuracy score. Feature importance using the gain method had more zip codes. Whereas feature importance using cover method had more GEOID and feature importance using weight had more numerical columns.

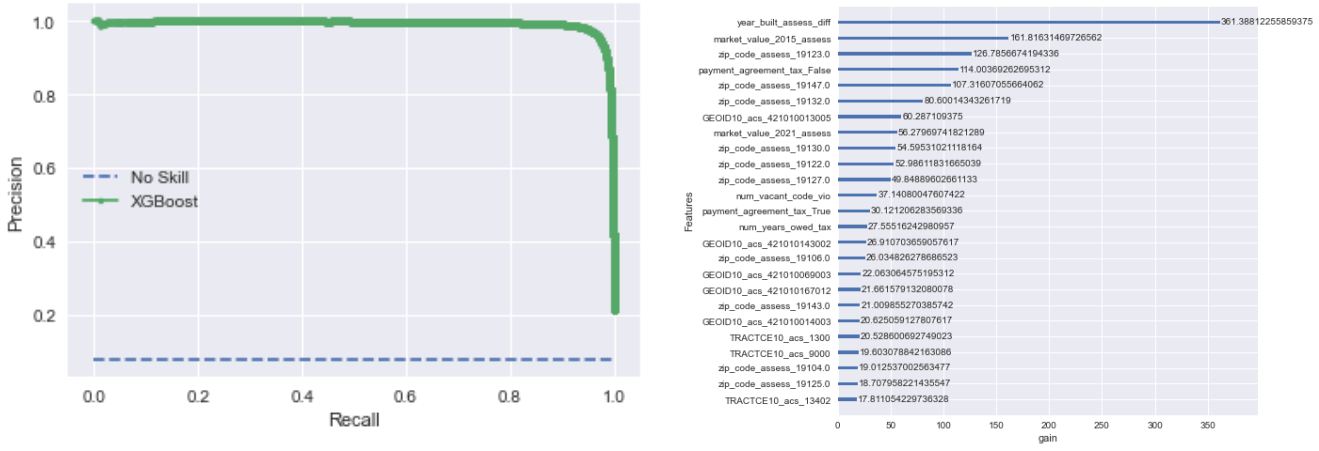


Figure 11 & 12. XGBoost Precision Recall Graph and Feature importance from XGBoost calculated by gain.

A. Srivastava
P. Ball
R. Carissa

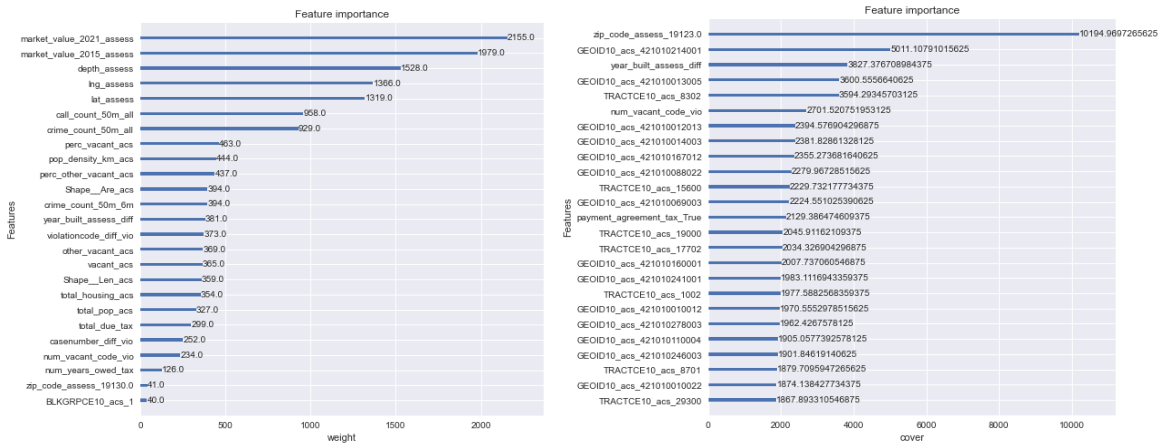


Figure 13 & 14. Feature importance XGBoost calculated by weight and Feature importance from XGBoost calculated by cover.

Deployment

The final model was deployed on Streamlit which can easily turn python scripts into shareable web apps. The app is simple and interactive. It lets the user select an address from the list of available addresses in Philadelphia and automatically predicts if the address is vacant or not. The app also shows the most important features from the model used. The app could use more real-time data to let users know the vacancy status of the addresses they are interested in.

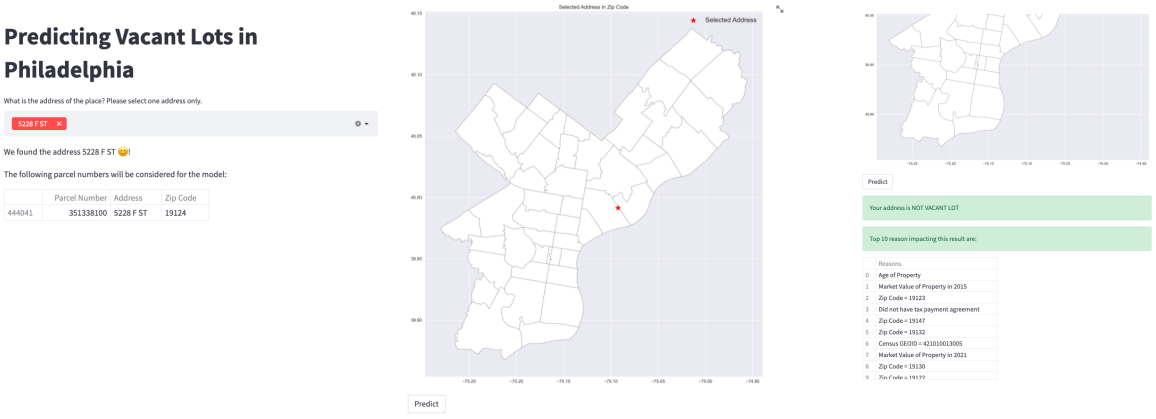


Figure 15. Screenshots from the deployment. Video of deployment can also be found here

V. Discussion

Since many datasets are insufficient or contain many null values, we use various datasets to get as close as possible to the true values of each parameter. There are, however, several challenges in using different datasets, which may adversely affect the quality of the output of the

model we are trying to deploy.

Working with several datasets not only took a long time due to the size of the compiled data but also because it adds more columns that we need to dissect in preparation for the preprocessing and extraction process. Understanding variables in different datasets is more challenging due to the inconsistency and complexity of the presented data. As an example, we went through different iterations of `parcel_number` to find that `opa_number` represents the same data.

With a large dataset, this slows down the processing time and makes it more difficult to draw clear conclusions by looking at the plotted maps. For example, before we arrived at the 50-meter average radius between houses in Philadelphia, we tried running higher numbers (up to 500-meters) but our system simply does not have the capacity of that scale to run the full analysis. We ran into similar issues while trying to hyperparameter tuning some of the models like XGBoost. With a large amount of data and the time constraint we have on this project, we decided to omit other relevant datasets, such as race, age, and income group.

The project can be expanded and improved further by including other census datasets like education, poverty, household income, race, poverty status. Data from Open Street Map can also be incorporated into the research. Further models can be tested including models on neural network. We can also use similar methodology to predict vacant lots in other cities.

A. Srivastava
P. Ball
R. Carissa

References

Bass, Margaret, et al. *Vacant Properties: The True Cost to Communities*. National Vacant Properties Campaign, Aug. 2005, https://community-wealth.org/content/vacant-properties-true-cost-communities.

Branas, Charles C, et al. “Vacant Properties and Violence in Neighborhoods.” *International Scholarly Research Network*, Department of Biostatistics and Epidemiology, University of Pennsylvania, 10 Sept. 2012, https://repository.upenn.edu/cgi/viewcontent.cgi?article=1004&context=cml_papers.

Branas, Charles C., et al. “Adding Windows to Vacant Houses and Clearing Vacant Lots Reduces Gun Violence, Saves Money.” *Pennmedicine.org*, Penn Medicine, 13 Oct. 2016, https://www.pennmedicine.org/news/news-releases/2016/october/adding-windows-to-vacant-house.

Chen, James. “Sheriff's Sale.” *Investopedia*, Investopedia, 19 May 2021, https://www.investopedia.com/terms/s/sheriff-sales.asp.

Chowdhury, Md Towhidul Absar. “A Machine Learning Approach on Providing Recommendations for the Vacant Lot Problem.” *RIT Scholar Works*, Rochester Institute of Technology, May 2017, https://scholarworks.rit.edu/theses/9428/.

“Code Compliance.” *Welcome to Burlingame, California*, https://www.burlingame.org/departments/building/code_compliance.php.

Esposito, Nic. “Philly's Vacant Lot Problem Could Be a Community Opportunity, Rather than a City LiabilityNic.” *Grid Magazine*, Grid Magazine, 18 Sept. 2020, https://www.gridphilly.com/blog-home/2020/9/17/phillys-vacant-lot-problem-could-be-a-community-opportunity.

Favre, Lauren. “Cities With the Highest Intensity of Gentrification.” *U.S. News*, U.S. News, 2 Aug. 2019, https://www.usnews.com/news/cities/slideshows/cities-with-the-highest-percentage-of-gentrified-neighborhoods.

“Foreclosure and Disposition of Tax-Delinquent Properties.” *Local Housing Solutions*, 8 Feb. 2022, https://localhousingsolutions.org/housing-policy-library/foreclosure-and-disposition-of-tax-delinquent-properties/.

Gobster, Paul H., et al. “Measuring Landscape Change, Lot by Lot: Greening Activity in Response to a Vacant Land Reuse Program.” *Landscape and Urban Planning*, Elsevier, 31 Dec. 2019, https://www.sciencedirect.com/science/article/pii/S016920461930917X.

Greenberg, Miriam, and Susie Smith. “Environmental Gentrification.” *Critical Sustainabilities*, 17

Aug. 2020, https://critical-sustainabilities.ucsc.edu/environmental-gentrification/#:~:text=%22Environmental%20gentrification%E2%80%9D%20is%20the%20process,and%20displace%20low%2Dincome%20residents.&text=Much%20discussion%20of%20gentrification%20has,shifts%2C%20rent%20theories%2C%20etc.

Loesch, Maggie. “Greening Vacant Lots: Low Cost, Big Effect in Philly.” *Shelterforce*, 22 July 2020, https://shelterforce.org/2018/11/13/greening-vacant-lots-low-cost-big-effect-in-philly/.

Mallach, Alan. *The Empty House Next Door: Understanding and Reducing Vacancy and Hypervacancy in the United States*. Lincoln Institute of Land Policy, 1 Jan. 2018, https://www.researchgate.net/publication/335110328_What_Makes_Mixed-Use_Development_Economically_Desirable_httpswwwlincolninstedusitesdefaultfilespubfilesshen_wp20qs1pdf.

Nassauer, Joan Iverson, and Julia Raskin. “Urban Vacancy and Land Use Legacies: A Frontier for Urban Ecological Research, Design, and Planning.” *Landscape and Urban Planning*, Elsevier, 2 Mar. 2014, https://www.sciencedirect.com/science/article/abs/pii/S0169204614000309.

Pearsall, Hamil, et al. “The Contested Nature of Vacant Land in Philadelphia and Approaches for Resolving Competing Objectives for Redevelopment.” *Cities*, Pergamon, 23 May 2013, https://www.sciencedirect.com/science/article/abs/pii/S0264275113000498.

“Philadelphia Population.” *World Population*, https://www.populationu.com/cities/philadelphia-population.

“Real Estate Tax.” *Phila.gov*, https://www.phila.gov/revenue/realestatetax/.

Reyes, Eduardo B, et al. “Early Detection of Properties at Risk ... - Dssgfellowship.org.” *Center for Data Science and Public Policy, University of Chicago*, The Department of Buildings & Inspections, City of Cincinnati, 2016, https://www.dssgfellowship.org/wp-content/uploads/2016/10/34_blancas.pdf.

Rigolon, Alessandro, et al. *What Predicts the Demand and Sale of Vacant Public Properties? Urban Greening and Gentrification in Chicago*. Pergamon, 1 Oct. 2020, https://www.sciencedirect.com/science/article/pii/S0264275120312968.

Schilling, Joseph, and Kimberley Hodgson. “Philadelphia’s Vacant Property Journey: Fostering Collaborative Alliances with Converging Policy Reforms.” *Vacantpropertyresearch.com* , Vacant Property Research Network, 2013, https://vacantpropertyresearch.com/wp-content/uploads/2013/09/Philly-Layout_V5.pdf.

“Sequestration.” *European Environment Agency*, 14 Feb. 2017, https://www.eea.europa.eu/help/glossary/gemet-environmental-thesaurus/sequestration.

Shlay, Anne B., and Gordon Whitman. *Blight Free Philadelphia: A Public-Private Strategy to Create and Enhance Neighborhood Value*. Temple University Center for Public Policy., 2001.

"Vacant and Abandoned Properties: Turning Liabilities into Assets: HUD USER." *Vacant and Abandoned Properties: Turning Liabilities Into Assets | HUD USER*, 2014, https://www.huduser.gov/portal/periodicals/em/winter14/highlight1.html.

"Vacant Land Management in Philadelphia: The Cost of the Current System and the Benefits of Reform." *Econsult Solutions, Inc.*, 23 Sept. 2010, https://econsultsolutions.com/vacant-land-management-in-philadelphia-the-cost-of-the-current-system-and-the-benefits-of-the-reform/#:~:text=News%20%2B%20Insights-,Vacant%20Land%20Management%20in%20Philadelphia%3A%20The%20Cost%20of%20the%20Current,and%20the%20Benefits%20of%20Reform&text=Econsult%20estimated%20that%20the%20current,million%20in%20lost%20property%20taxes.

"Vacant Lot Program: Programs and Initiatives." *City of Philadelphia*, https://www.phila.gov/programs/vacant-lot-program/.

Wachter, Susan M., and Kevin C. Gillen. "Public Investment Strategies: How They Matter for Neighborhoods in Philadelphia." *ResearchGate*, The Wharton School of Business of the University of Pennsylvania, 1 Jan. 2005, https://www.researchgate.net/publication/266878799_Public_Investment_Strategies_How_They_Matter_for_Neighborhoods_in_Philadelphia.

"What Can I Do with My Vacant Lot in Philadelphia?" *Brotherly Love Real Estate*, 16 Apr. 2021, https://brotherlyloveproperties.com/what-can-i-do-with-my-vacant-lot-in-philadelphia/.