Massachusetts Housing Partnership
CS 506: Final Report

Camilla Belamarich
Lingxiao Yuan
Alex Luan
Asif Rahman

Residential Housing Units and Density in Massachusetts

Motivation and Project Description

Currently, there is no accurate or available data on the number of housing units in Massachusetts. Without this knowledge, it is hard to supply the communities of Massachusetts affordable housing. Previous to this project, the Massachusetts Housing Partnership (MHP) team developed a process to calculate the residential density of areas near every MBTA rail station using the statewide assessors parcel and statewide emergency 911 data. This project was called TODEX (Transit-Oriented Development Explorer). Our goal is to use the same two data sets and calculate the residential density for specific areas of land across the entire state of Massachusetts. The methodology developed in TODEX was useful, but extremely time consuming to replicate and difficult to scale. We aim to accomplish these goals in a few steps that will be easily reproducible for the MHP team when the semester ends. We will first scale down the data sets into individual municipalities, apply our pre-processing methodology to only include residential units, merge the two data sets, and calculate residential density. In order to make this useful for the MHP team, we thoroughly documented our code and process.

Data Description

For this project, we worked with the statewide assessor parcel data and statewide address data. The Standardized Assessors' Parcels data is publicly available on the Mass.gov website and is provided by MassGIS (Bureau of Geographic Information). It contains parcel boundaries and database information from each community's assessor. The statewide parcels database is a geodatabase, meaning it contains geographic information. Specifically, it has information on the parcel and its location, a LOC_ID, number of units, USE_CODE, and geometry of the parcel. The LOC_ID is an extremely useful feature of the data set that will be used to calculate residential density. Additionally, the number of units is a feature that will be useful to calculate residential density and report the number of units for each parcel. The Use Code is an alphanumeric value that assigns a type to each parcel. Since we are only focusing on residential units, this feature is useful for filtering out non-residential units. The geometry consists of polygon coordinates of the parcel, which is a geographic shape made up of points. This feature is useful when merging the two data sets and matching the addresses to corresponding parcels.

The second data set, Master Address Data, was also provided by MassGIS and can be found on the Mass.gov website. Address data from this data set was compiled from the state 911 Department to provide the most comprehensive list of standardized addresses for cities and towns throughout the state of Massachusetts. This is also a geodatabase and contains information for each address and geometry. The geometry feature in this data is the actual geographic points, while the geometry in the parcel data set is the coordinates of the shape of the parcel.

Data Download and Processing

After downloading statewide parcel and address data sets from the Massachusetts government website, we attempted to read in the geodatabases using geopandas. However, we ran into multiple problems with file size, which made it difficult to start preprocessing. Additionally, our laptops did not have enough RAM to be working with files this large. To resolve this problem, we converted the necessary files to feather format. This essential step allowed us to access the files and continue with our preprocessing. Since both data sets are geodatabases, using geopandas was the appropriate choice to read in and process the data.

To process the data sets, we first observed each feature and learned what they were used for. Since our goal was to calculate residential density and report the number of units for each parcel, we looked at the behavior of the data for LOC_ID, Units, Use Code, and geometry. Both data sets contain data on all parcels across Massachusetts, both residential and non-residential. We learned that Use Codes that start with zero, one, or nine are considered residential. After removing all non-residential Use Codes in the statewide parcel, we observed invalid addresses. To our knowledge, these invalid addresses had the same Centroid Ids, but different number of units. Additionally, we removed these from the data set. In order to merge the two data sets and start our analysis, it was necessary to first preprocess the data. This step made it less computationally intensive to merge the data sets and allowed us to only focus on the residential parcels.

Key Question 1: How to merge state-wide address data and state-wide parcel data?

Once the preprocessing was done, we were able to start the merging process of the statewide parcel and address data. This step was the most time consuming due to the discrepancies between the data sets of the reported number of units for some parcels. Although we filtered out non-residential parcels based on Use Code, there were still many instances of non-residential parcels. We started assigning values to parcels based on the description of the Use Code. For example, “2 Family House” was assigned two units. However, there were some Use Codes where we could not apply this rule to. In order to merge the data sets, we used geopandas “sjoin” feature, which is a spatial join that combines data sets based on a relationship. In this case, the relationship is the parcel polygons and the address points. This merge will assign addresses to their corresponding parcels. Subsequently, we counted the number of addresses in each parcel.

In order to accurately assign a unit value to Use Code, it was necessary for us to create a set of decision rules for instances when we were not able to assign unit values to parcels based on Use Codes. This helped us verify the data set after merging and determine which unit count was more accurate to use for certain parcels between the two data sets. We labelled instances of when we could assign a unit value to a Use Code, “Unit Countable”. We then labelled instances of when we could not assign a unit value to a Use Code, “Not Unit Countable”. For Not Unit Countable parcels, we decided that using the statewide address data was more accurate than the statewide parcel data. For example, the parcel data set reported only one unit for an entire apartment building because this data is based on reported taxes of that parcel. However, this is inaccurate because apartment buildings have more than one unit. The Unit Countable and Not Unit Countable decision rules are displayed in Table 1.

Table 1. Unit countable Use Codes.

Unit	Countable	
USE CODE	TYPE	PROOF
0101	Single Family	proof_0101.txt
0102	Condo	proof_0102.txt
0104	Two Family	proof_0104.txt
0105	Three Family	proof_0105.txt
010E	Two Family	proof_010E.txt
010G	Two Family	proof_010G.txt
010H	Two Family	proof_010H.txt
010I	Single Family	proof_010I .txt
010M	Single Family	proof_010M.txt
101	Single Family	Official documents
102	Condo	Official documents
104	Two Family	Official documents
105	Three Family	Official documents

We developed a process in which to apply these decision rules to each parcel across Massachusetts. For Use Codes that were determined to be Unit Countable, the number of reported units matched the number of addresses, we would use the counting results due to the correct matching of the assigned unit value to reported number of units in the addresses data. Conversely, if the number of assigned unit values does not match the reported number of addresses, we would mark these as anomalies.

- “2 Family Home” → 2 addresses reported at this address = Unit Countable
- “2 Family Home” → 5 addresses reported at this address = Anomaly

Anomaly detection was an essential step in our data mining process due to the many instances of discrepancies between the two data sets. In order to verify our predictions of unit count of these anomalies, we performed linear regression to make assumptions that minimize the distance between the data points and the regression line (Fig. 1). After predicting the unit counts for the anomalies, we plotted our results on an interactive map, which indicates red houses for anomalies and blue houses for non-anomaly parcels.

Table 2. Not Countable Use Codes.

Not USE CODE	Unit TYPE	Countable PROOF
0103	Mobile home	proof_0103.txt
0107	Not Clear	proof_0107.txt
0108	Not Clear	proof_0108.txt
0109	Multi House	proof_0109.txt
010C	Not Clear	proof_010C.txt
010F	Not Clear	proof_010F.txt
010J	Not Clear	proof_010J.txt
010Z	Not Clear	proof_010Z.txt
011	Apartments	proof_011.txt
012	Not Clear	proof_012.txt
013	Not Clear	proof_013.txt
014	Not Clear	proof_014.txt
015	Not Clear	proof_015.txt
016	Not Clear, Mainly apts	proof_016.txt
017	Not Clear, Mainly apts	proof_017.txt
018	Not Clear, Mainly apts	proof_018.txt
019	Not Clear, Mainly condos	proof_019.txt
021	Not Clear	proof_021.txt
031	Not Clear	proof_031.txt
041	Not Clear, Mainly apts	proof_041.txt
051	Not Clear	proof_051.txt
061	Not Clear	proof_061.txt
071	Not Clear, Mainly apts	proof_071.txt
081	Not Clear, Mainly apts	proof_081.txt
091	Not Clear, Mainly two fam	proof_091.txt
103	Mobile Home	Official documents
107	Not Clear	Official documents
108	Not Clear	Official documents
109	Multi House	Official documents
11X	Apartments	Official documents
12X	Non-Transient Group Quarters	Official documents
945	Not Clear	proof_945.txt
959	Not Clear	proof_959.txt
970	Not Clear	proof_970.txt

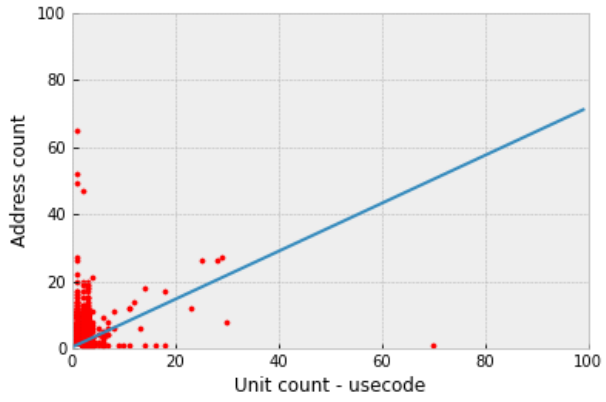

Figure 1. Linear Regression Plot of Unit Count. The results of Linear Regression analysis to make assumptions of unit count on parcels determined to be anomalies.

Key Question 2: What is the residential data of each LOC_ID?

After predicting unit counts for all the anomalies in our merged data set based on our assumptions, we calculated residential density per LOC_ID. Calculating residential density is one of the main goals of this project. This was done by dividing the area, which is a feature in the parcel data set, by how many units were counted in a parcel per LOC_ID. The Area feature in the parcel data set was measured in square feet. We converted square feet to acres and reported the residential density in our final data set.

Key Question 3: How can we make this reproducible and updatable for MHP after this project?

The most important step we took to make this process reproducible to MHP and future spark projects was to heavily and thoroughly document our code. This step allows for future modifications and additions. The data sets are constantly being updated, so it was important to include this step in our process. In total, our whole process takes thirty minutes from start to finish. We determined that this amount of time would be sufficient enough for MHP to insert a new datafile directly into our code when updated information is available to them.

Final Data Set Description

The final data set consists of the merged statewide parcel and address data with our added modifications, calculations, and assumptions. LOC_ID, Use Code, Geometry, and City remained the same from the original data sets. Style Descriptions refers to the Use Code description we used to determine residential and non-residential parcels. Count Use Code is the number of units counted based on the Use Code. This is similar to what we have done with Unit Countable Use Codes. Area_Sqfeet refers to the area of the parcel, which retains information from the original parcel data set. Add_count refers to the number of addresses in each parcel. The assumption feature refers to the unit count predictions we made using unit count collected by Use Code and address count. This was the final assumption we made based on our decision rules. Is_anomaly indicates whether the parcel is an anomaly or not. Density_Sqfeet refers, Density_Sqmeter, and Density_Acre refers to the residentials density calculator per square foot, square meter, and acre, respectively.

We also created an interactive map for all parcels across Massachusetts (Fig. 2). This allows users to hover over parcels and learn about the style, unit count, and density. This pop-up shows the LOC_ID, style of the parcel, Use Code, address count, unit count, our final assumption, and density. In the future when more accurate data is available, parcels that are labelled red, anomalies, can be updated. We plotted this map using Python's Folium package. This is where the geometry feature in the final data set was extremely useful. Our process for plotting these maps is designed in which the user can plot by city if they are interested in re-analyzing or re-updating one city in specific. Having a visualization tool for address data can be extremely helpful when assessing certain cities to see if the data needs to be updated.

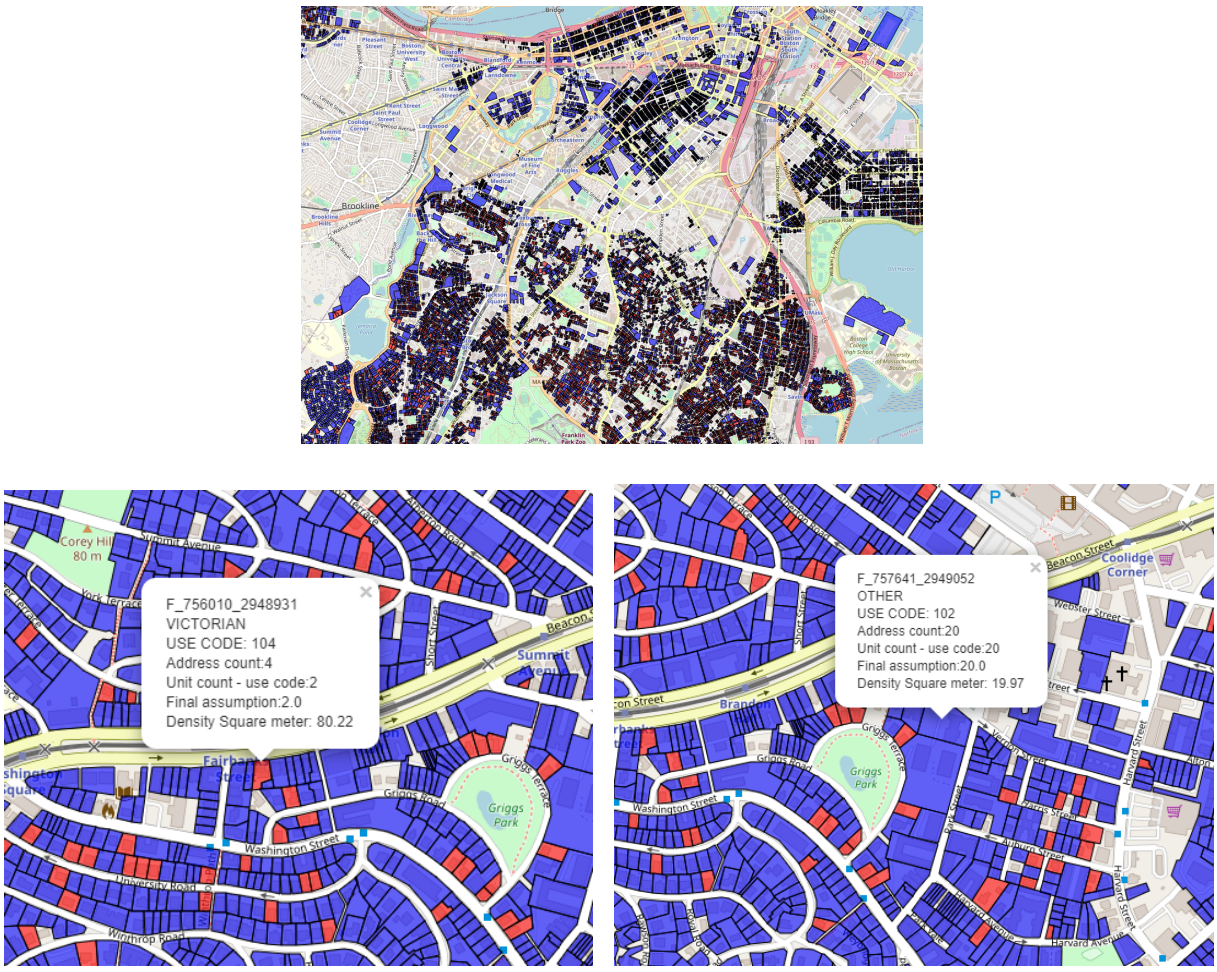

Figure 2. Interactive Map to View Unit Count and Residential Density. Top image shows the town of Brookline and residential units. Red parcels represent the anomalies we detected in our analysis and the blue parcels represent non-anomalies.

Limitation and Challenges

We ran into many challenges using the statewide parcel and address data set. Both data sets were determined to be inaccurate with regards to the reported number of units, which is the reason we had to develop a process to decide which unit count to use for certain parcel types. Additionally, the magnitude of the data sets made this an even bigger challenge for us. When we scaled down the data set for evaluation to just Quincy, there were many anomalies and discrepancies between the data sets. Since this was a challenge at a smaller scale, we knew we had to most accurately predict the unit counts for anomalies. So when it was time to apply our process to the entire state of Massachusetts, we minimized the potential for error.

Preprocessing and merging the data sets was time consuming and the process could be more efficient. We used “sjoin” to merge the data sets, which was efficient; however, cleaning the data beforehand took up most of the time. Cleaning the data of duplicates, inavlid addresses,

and non residential parcels was an essential step before merging, and since the data sets were millions of lines long, this was extremely time consuming. We used goepandas/pandas built-in function to group data by specific features, like Use Code or Town ID, to observe unit count and duplicates in the data sets. Although the cleaning and merging steps of our process took away from time we could have used to analyze the data more, both data sets were complex and unstandardized. Furthermore, spending the amount of time we did on these steps was necessary for a more accurate final product.

Table. 3(a) Unit count of by Address Dataset for LOC_ID = F_783701_2920473

	CENTROID_ID	STREET_NAME	FULL_NUMBER_STANDARDIZED	UNIT	geometry	index_right	LOC_ID
2153027	M_238871_890163	GRANGER STREET	82	2 LEFT	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153028	M_238871_890163	GRANGER STREET	82	1 RIGHT	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153429	M_238871_890163	GRANGER STREET	82	None	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153430	M_238871_890163	GRANGER STREET	82	1	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153431	M_238871_890163	GRANGER STREET	82	4	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153432	M_238871_890163	GRANGER STREET	82	1 LEFT	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153433	M_238871_890163	GRANGER STREET	82	3	POINT (238871.062 890163.422)	97.0	F_783701_2920473
2153434	M_238871_890163	GRANGER STREET	82	2 RIGHT	POINT (238871.062 890163.422)	97.0	F_783701_2920473

Table. 3(b) Unit count by Assessor Parcel Dataset for LOC_ID = F_783701_2920473

	LOC_ID	TOWN_ID	USE_CODE	ADDR_NUM	SITE_ADDR	STYLE	SHAPE_AREA	geometry	parcel_count
13415	F_783701_2920473	243	1110	82	82 GRANGER ST	4 Fam	774.707227	POLYGON ((238868.485 890182.982, 238892.317 89...	4.0

Table 3 shows one example of the Address dataset having duplicates. Based on the parcel data we know that this building is a four family house, which should have four units(1,2,3,4). However the Address data have recorded 4 duplicates(‘1 left’, ‘1 right’, ‘2 right’ and ‘none’) We later cleaned the Address data by removing those invalid ‘none’ units. However there are still other types of duplicates in the Address dataset that need to be cleaned/removed from the dataset in the future work.

Mixed-use parcels also caused pre-processing issues. Mixed-use parcels are buildings, for example, that have a commercial unit on the ground floor and residential units above it. This is extremely common in more urban areas, and especially common in Boston. Since these Use Codes are not standardized, it was difficult for us to form assumptions about the final unit count. We decided it was best to use the address data unit count for parcels like this.

An additional limitation with Use Codes is that Use Codes starting with nine were considered residential; however, there were many instances of non residential parcels. For example, there were many instances of churches, schools, and dormitories. There were also special instances of charitable housing parcels, which we were hesitant to consider as non residential parcels since they were described as “housing”. The majority of these parcels are tax-exempt, meaning that there would be no record of them in the parcel data set, but they would appear in the address data set, which does not describe the type of address it is.

Lastly, Use Codes were not standardized and not accurate if we compared them to different towns. To overcome this standardization problem, we combined the Use Codes to each town ID. There was around two-thousand Use Codes, which we had to go manually through due to unstandardized descriptions of each code. For example, Use Code 104 had the description "Two Family Residential" and Use Code 104A had the description "Other". There were many cases of this happening and the corresponding number of units differed also. Manually going through them was time consuming; however, we could have used NLP to perform classification using Use Code Descriptions. Since the descriptions of the Use Codes were also not standardized, it is necessary to use a more complex classification method to overcome these challenges. Due to the time-constraint of the semester, we were unable to perform this classification algorithm.

Conclusions and Next Steps

Overall, we successfully preprocessed and merged the two statewide data sets, reported the number of units for each parcel, calculated the residential density, and provided an interactive map for visualization. We answered our three key questions and documented our code for future use. Our final dataset and map will be useful for MHP to have a better understanding of the statewide housing status and provide the communities of Massachusetts with affordable housing. We hope with our documented process, they will be able to update the datasets when new data is provided to them and modify our process if inaccuracies are detected in the future. Efficient next steps, possibly for future BU Spark projects, would be to apply a more complex classification model to better predict unit count for parcels that are determined to be anomalies based on the Use Code and Use Code description. Additionally, optimizing the preprocessing and merging steps based on our initial decision rules would lead to a more efficient analysis of the data. Given that there was not an available or accurate tool to observe statewide housing and available housing units, our project produces a strong foundation for a more concrete Massachusetts housing database.