Project Report

By Martin Yim and Sean Zhang

Introduction/Motivation

How do people make decisions in urban environments regarding time, money, and location? Where people work and live, and where they go in between is often a function of people's preferences and constraints. Workplace location is usually not flexible, but people do have choice in where they live, and how much time or money they spend on a home and during a commute. We aimed to examine some of the factors and choices people confront when looking for housing. The analyses that follow are a brief venture into what could be a large repository of interesting problems about living in urban areas, and how to optimize for one's own needs.

Data Review

Our data primarily came from the City of Boston Open Data portal. The Boston Open Data portal contains datasets on a variety of different aspects of the city's pulse. We worked with the MBTA API, Google Maps API, a list of Boston Zip Codes, and the Property Assessment Data from 2014 to do our analysis.

Other datasets that we gathered in the initial stages of the project were: parking ticket data, snow parking data, and vehicle excise tax data. We initially were hoping we could use this data to analyze transportation in Boston: where are the gaps in public transportation services that are being covered by private services (eg: Uber, Taxis), and could public transportation fill those gaps. Due to lack of data on private transportation patterns, we eventually decided to forgo this idea for our housing and commuting analysis.

Using the MBTA API, we gathered data about each subway (T) station and bus stop. For each station/stop we gathered information relevant to our analyses: the stop id, longitude and latitude, and the stop's respective route. One of the interesting challenges we had with the MBTA API was that when we initially got all the data from the API, we found that we had almost 8,000 entries. This seemed very illogical; how could there possibly be 8,000 MBTA stops throughout all of Boston? After further inspection, we realized that many data points of the MBTA API overlapped (i.e. the B, C, and D on the Green line all overlap past Kenmore). Once we had aggregated all of our data for project 1, it was time for our statistical analysis and constraint satisfaction problem.

Both the statistical analysis and constraint satisfaction problem required information on real estate. As such, we primarily used property assessment information from 2014 in the City of Boston Open Data portal. From this dataset we had information about most properties within Boston. Some examples of the important information the property assessment (2014) dataset provided were: assessed property value, location (longitude and latitutde), zipcode, and building style.

Constraint Satisfaction

The constraint satisfaction problem we solved was an interesting, yet challenging, question to answer.

Let's assume that we have: the cost of a property we currently live in, the walking time to get to the nearest public transit station, and the commute time via public transit. Given this information, we would like to find other potential places to live within the same zipcode such that one would be indifferent to moving between the current home location and the potential new locations based on those location's respective costs, walking times, and commute times.

Our constraints are the costs, walking times, and commute times. We want to answer the question: does there exist a place(s) in which we are indifferent compared to our current location and its commuting attributes? This kind of constraint satisfaction problem could actually be very applicable for various reasons. For example, say I live in Allston and take the T to commute to work at Kendall Square. Suddenly, city planners have decided to renovate the entire block that I live on in Allston. They could offer a suitable replacement in Fenway with a different walk and commute time, but with confidence that I would be indifferent in terms of my overall trade-offs for the times and costs needed at these two locations. Our research would be able to check and verify the question: does there exist a place I would be indifferent towards living?

To answer this question, we first decided to hard code a company location as a parameter: Kendall Square. Then, we needed to gather extra data; namely, finding the closest MBTA stops for a given residential property. To find this, we queried the Google Maps API to get the nearest public transit stops within 1km of a residential property. With this data, we based our walk time and commute time on general averages: a person can walk a kilometer in approximately 10 minutes and public transportation can travel a kilometer in 6 minutes (with traffic).

Once we had the walk times and commute times, we combined these with the assessed property value and passed them through the z3 library, which helped us determine whether our constraint was satisfiable or not.

Our equation was of the form:

$$(w*w_i + t*t_i + r*r_i) == 1$$

For all i within a region (zipcode), where w_i, t_i, and r_i, are the walk times, transit times, and residential costs, respectively. We then hand this equation to z3 to solve for w, t, and r.

However, we found that this was too strict of a requirement. Within any given zipcode, there were thousands of properties, walk times, and zip codes. Finding a combination that strictly equated to 1 was close to impossible. As such, we loosened the constraints just slightly:

$$(w*w_i + t*t_i + r*r_i) <= 1$$

$$(w*w_i + t*t_i + r*r_i) > 0.99$$

This allowed us to find at least five zip code regions where a person would satisfy the indifference constraint.

Statistical Analysis

We built a multiple regression model to predict assessed property value based on less common factors, in conjunction with more traditional factors.

Factors:

- Number of rooms in property
- Number of bedrooms in property
- Number of full bathrooms in property
- Building area
- Year property was built

Dependent Variable:

- Assessed property value

Our regression model had an R-squared of 0.635, indicating that roughly 64% of the variance in our dependent variable (assessed property value) could be explained with the five factors we used. Although our model has a fairly high R-squared it is not sufficient on its own to be a reliable predictor of assessed property value. However, the current model can at least help us understand the relationship between the respective factors and the assessed property value.

The relationships between our factors and assessed property value were surprising; our model construction may have contributed to the non-intuitive results. Our results would suggest that each full bathroom contributes roughly $88,000 to the assessed property value, while at the same time, additional rooms subtract $43,000 from the assessed property value. Also, each additional bedroom subtracts $34,000 from the assessed property value. Each extra square foot of building area contributes $243 to the assessed property value, and each year closer to 2014 (highest year available) makes the property worth $48 more. All of these effects have non-zero inclusive confidence intervals, suggesting they are all significant factors with 95% confidence.

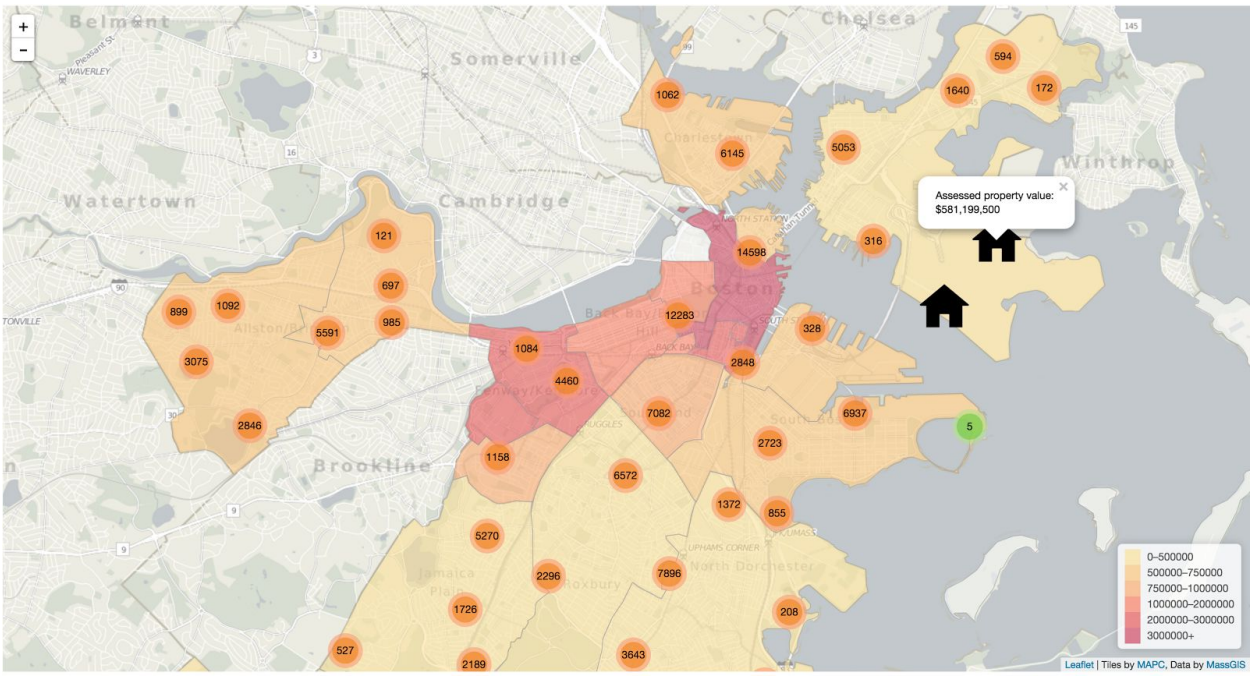
Although our model gives unexpected and strange results, this can be explained. It is likely that our unconventional choices in variables have enabled omitted variable bias to create massive distortions within the coefficients (effect size) of our factors onto our dependent variable. For example, it is likely that location has high explanatory power with regards to assessed property value, and yet, we did not include a location variable in our model. This likely makes up a significant portion of the 36% of variance our model cannot explain (1 - R-squared) . Multicollinearity may have also posed a problem with our model; number of bedrooms, number of rooms, and number of bathrooms are all likely to be highly correlated with each other. This multicollinearity between variables can disrupt the model and create inaccuracies. A final possible explanation for the strange coefficients is sample selection. A significant portion of the

dataset was lost due to missing values during the modeling. If the missing values are systematic in nature, they could have a significant impact on the model's relation to reality. Roughly 62,000 observations in the dataset were included, but almost as many were excluded due to missing data for the relevant factors.

Future analysis should emphasize better sample selection and better feature selection/engineering and diagnostics to avoid omitted variables bias, sampling biases, and multicollinarity.

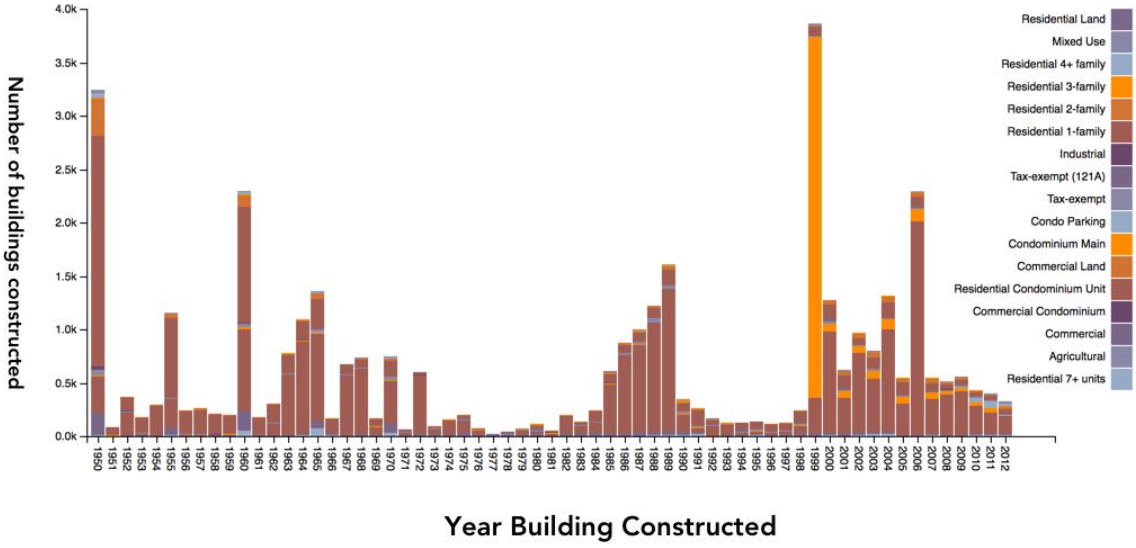
Visualizations

Once we aggregated all of our data, we thought it would be useful to map out the average property valuation for each neighborhood within Boston. As such, we used Leaflet to generate this visualization:



We first converted all of the residential property data into GeoJSON. Then, using Leaflet's native clustering library, we combined nearby locations to better visualize how many properties were within a given region. The user can then interact with the map and zoom in or zoom out to visualize bigger or smaller clusters. Once the user zooms in close enough such that there are no more clusters, all of the relevant properties are denoted with a house icon.

Our second visualization is a stacked bar chart of the number/types of properties over a 50 year time period from 1950 to 2012:



We thought that this visualization would be interesting to see what kinds of properties were being built over time. We can see that, for the most part, there is a general trend of residential 1 family properties being built every year. The trend follows the natural rise and fall of the housing market. Every 8-10 years, the housing market fluctuates as the economic situation fluctuates. We thought that this was a relevant visualization because while researching our residential optimization problem, it's useful to have a sense of the market trend and how many houses are being constructed. The spike in 1990 stands out in particular.

We were curious as to why there was a sudden rise in residential 3 family properties, and we hypothesize that this is possibly due to a massive increase in market price for housing in Boston between 1995 and 1999. During that period housing prices increased by roughly 35%[1]. The price increases were due to a constrained supply of housing. We hypothesize that the massive 1999 housing construction boom was a result of an effort by developers to capitalize on high housing prices in Boston, and what was perceived as a hot market for real estate, and was also a market correction allocating resources to a market lacking in housing supply.

Concluding Remarks

Overall, we found interesting results from both our constraint satisfaction problem and statistical analysis. However, the way that we constructed the constraint satisfaction problem was counter-intuitive. To make our constraint satisfaction problem nontrivial, we effectively found *people* with preferences towards walk time, commute time, and residential cost, as opposed to

[1] http://www.tbf.org/tbf/51/~/media/E3ECCE393A8F4FDC83C3B95459AC772D.pdf

our original goal of finding *places* given the constraints. Our statistical analysis also presented us with some interesting traits, but ultimately it may not be able to tell us much. There were various intercorrelated factors that may have caused the way we viewed the data to be completely skewed.

Future Work

Given the opportunity to pursue this further, one of the things that we could work on is coming up with a way to better construct our constraint satisfaction problem. We already have the walk time data, the commute time data, and the residential property costs. Ideally, we could try to expand the constraints into a variety of more options, i.e. Living close to a school because I'm a parent, living close to family, etc. The applications of this constraint satisfaction problem are endless, and ideally if we could solve it, we could genuinely help people find their next home and still be content.

Future statistical analysis or modeling efforts would be best served by a greater focus on feature selection. Adding factors such as location would help to increase the predictive power of the model, and hence tell us more about housing prices in Boston.

Though the results of our work may be inconclusive, it sets up the groundwork for future exploration in urban, human-centered design, and how to optimize our lifestyles with respect to our urban lives.