Daren McCulley

CS 591 Project Proposal

Abstract

Cities are better engines for social, economic, and cultural development when they are hosts to residents of all income levels. Even if you reject that premise, a simpler case can be made for the idea that cities require a workforce of all income levels in order to properly function. It follows then, that it is in a city’s best interest to allow residents of all income levels to find housing within city limits or provide cost-effective and expedient transportation options in and out of the city from affordable suburbs. Unfortunately, in many cities parcels which once housed three or four families are being gutted and redeveloped to make room for one wealthy family displacing a large and crucial segment of the population. Neighborhoods which once brought people together from all walks of life are transforming into walled gardens for those with the ability to afford the price of admission.

I am not an urban planner, social scientist, or political scientist. I am a computer scientist. I am proposing mining an existing dataset, collected annually, along with what we know about the geography of a city to try and predict where the price of housing is likely to sharply rise in the coming years. I’ll leave it to those more qualified to decide what to do with that information. The complexity of gentrification makes the problem seem intractable. Even the word gentrification has no agreed upon definition, but are there two critical and more importantly *limited* resources that cities devote to fighting for a more socio-economically diverse population. The first is monetary, in the form of rent controls, subsidies, and tax policies. The second is political will, in the form of zoning, regulation, and protecting at-risk populations from the outsized influence of large developers and their wealthy customers.

Boston has been steadily climbing the list of cities featuring the widest gap between the ‘haves’ and ‘have nots’ for some time, which makes it a city worth studying. I’ve lived here for the last three years and witnessed the effects of this divide for myself, which provides a personal motivation for grappling with this issue. I enter into this research with no guarantee of success and reams of research papers dedicated to accomplishing the same goal with a far more complex model. Failure is an acceptable outcome, because I will learn that my model was too simple to be a good predictor, while honing my full-stack skillset.

Despite a high risk of failure, I believe what I’m doing is worth the effort, because my model does not rely on census data available once per decade, which is costly to gather and criticized for inaccurately capturing the reality of America’s poorest communities. Instead, I leverage data collected annually for the purpose of assessing property taxes, which is Boston’s largest source of revenue, and something they already endeavor to do accurately. I am deliberately trading model complexity for finer granularity along with low overhead. The best-selling book in the field of business research, *Good to Great*, demonstrated how massive and complex companies focused on a *single* metric by which to measure themselves and guide their decision making. I believe the same approach can and should be applied to more than running a business and I hope that this project produces a product, which is both novel, simple to understand, and with a little luck... accurate.

Description

In detail, what I am proposing is to gather the 2016 Boston Property Assessment data available here:

<https://data.cityofboston.gov/>; and to use the key-attribute “PID” in this data to scrape Boston’s Assessing Search Engine: <http://www.cityofboston.gov/assessing/search/>; effectively gathering up to 30 years’ worth of property valuation data along with other critical information like location and living area. The next step of my project involves tessellating the map of Boston and sorting each parcel into its host tile. Using MongoDB aggregation on location is possible at the DBMS level allowing me to calculate an average dollar value per square foot of living area for each year between 1985 and 2016.

Next, I will transform the tessellation into a graph, where an edge exists between two nodes if they share an edge in the tessellation. For this reason, I’ve elected to use a hexagonal tessellation, which has two nice properties:

1. Every tile’s center is equidistant
2. Allowing for six bidirectional edges will produce a well-connected graph and do a decent job modeling which communities share a border.

What is important to note is that a hexagon will only produce a node in the graph if it has a minimum threshold of living area within residentially coded parcels. For example, there will be no edge between the occupied tiles on the north side of the Boston Common and the south side, because of the unoccupied tiles within the Common itself. Large areas of tax exempt land like Harvard Business School or Longwood Medical Center may also have similar effects on the topography of this graph. A little known fact is that nearly half of Boston’s surface area is tax exempt, so I’ll be very interested to see what pressures if any that fact places on the market. Obviously, a key hyperparameter is the size of the tile. This is something I plan to experiment with, if time permits.

All of this work is aimed at generating a static state where each node in the graph has some starting value and the topology of the graph is at least modeled on the geography of Boston. My intention is to run a network flow algorithm, not unlike PageRank, which will produce a plot of inflow or outflow over time at each node. This is the information I intend to use as the *single* feature when training my machine learning model. I do not intend to classify nodes as “will gentrify” or “will not gentrify” because again, I am a computer scientist. Instead, I will try to predict what the dollar value per square foot will be in some node at some point in the near future based on using 30 years of known outcomes to train my model. If I can predict these values with some degree of accuracy then I would argue that I’ve at least given the decision makers who govern Boston another piece of information to help guide resource allocation.

[Obvious Hurdles: accounting for inflation; a lack of information beyond value earlier than present day; handling borders with neighboring cities like Brookline; selecting and tuning hyperparameters]

Collaborators

For now, I intend on doing the heavy lifting on gathering and processing the data alone. I’ll be asking two friends at Questrom seeking MBAs for advice on packaging the results for display and advice on use cases for this research as the project progresses. Their names are Monica Mondloch and Joseph Lucido, and I will cite any contributions they make.

Related Work

* <http://www.governing.com/gov-data/boston-gentrification-maps-demographic-data.html>
* <http://www.bostonredevelopmentauthority.org/housing>
* <http://www.kaynar-rohloff.com/papers/ACSA_Rohloff_Rohloff.pdf>
* <http://link.springer.com/article/10.1007/s101090200086>
* <http://www.sciencedirect.com/science/article/pii/S0198971506000718>
* <http://epn.sagepub.com/content/30/3/523.full.pdf+html>