Leo Kitchell

November 4, 2021

Senior Thesis – Working Draft

Gentrification Forecasting with Airbnb User Reviews

# Introduction

Gentrification is generally understood as a process of change whereby neighborhoods which have been historically under-capitalized experience an influx of new residents with high social capital. This demographic change corresponds with an increase in investment in the neighborhood, generally transforming the housing stock and character of the neighborhood. In theory, residents of gentrifying neighborhoods stand to benefit from increased retail and housing investment and the affiliated job opportunities and tax revenues. In practice, a gentrifying neighborhood’s original residents often find themselves forced out by a combination of passive market factors, namely increased rents, and active displacement mechanisms such as planned dilapidation, eviction, and reinvestment cycles. This widespread displacement of vulnerable populations is the principal concern of activists and is central gentrification’s contentiousness. Unfortunately, each of the three pillars of change (demographic, housing and retail stock, and cultural character) have characteristics which make quantifying their change difficult, and which consequently limits scholarship and effective policy intervention.

One of the principal data challenges is that detailed demographic data is infrequently collected, so researchers and policymakers are constantly fettered by data from the Census or the 5-year American Community Survey. This lag means that research must be conducted on sparse time series, limiting the research's power, while on the policy side the lag makes early identification of gentrification difficult, and dynamic monitoring impossible. Changes in the physical neighborhood real-estate composition are generally the easiest to detect, as home prices, building permits, and other measures of investment operate on shorter lag. The final pillar of change is also the most difficult to quantify. While the arrival of craft coffee shops, whole foods, and street art may be easily visible signs of cultural change for residents of a gentrifying neighborhood, this local knowledge does not get translated into traditional data sets. However, new data sources may solve some of these challenges.

Crowdsourced data from online platforms like Yelp, Zillow, and Airbnb are enabling researchers to investigate gentrification trends at a far more granular level and in real time. These data also supplement traditional metrics with novel features, such as user reviews and textual descriptions. These features represent an exciting frontier in gentrification research and may enable a better understanding of neighborhoods and their residents.

This paper utilizes 60 million text reviews from Airbnb listings in 28 U.S. cities to infer neighborhood characteristics which can be used to predict gentrification in real time. Our analysis is organized into 7 parts. Section 2 summarizes key empirical studies on gentrification with a focus on their specifications of gentrification and the geographies which are suitable for analysis. Section 3 introduces two recent studies which use low-lag data from Zillow and Airbnb, respectively, to predict measures of gentrification. Section 4 outlines our gentrification operationalization and two model specifications, with and without unstructured text data, which will be used to generate our results. Sections 5 and 6 present these results from the respective specifications. Finally, section 7 considers the implications of these findings and introduces an alternate model for near term gentrification prediction.

# Empirical Literature Review

The data challenges introduced in the previous section have hindered gentrification research, but a robust literature still exists, particularly when it comes to identifying areas that are ripe for gentrification. The term gentrification originates from the migration of London “gentry” into lower-income neighborhoods in the 1950’s and 1960’s (Glass 1964). Since the term’s coining, researchers have developed progressively more complicated methods of identifying gentrification, and considerable variation still exists in its operationalization. Among this backdrop of variation, resident income, education level, and race are commonly used measures of quantifying the first—demographic—pillar gentrification (Ding et al, Freeman 2005, Bates 2013, Goetz et al.). These variables generally describe the first pillar of gentrification, though significant changes in demographics may also signal a shift in the cultural character of a neighborhood. Although neighborhood demographic changes are necessary to establish gentrification, they are not sufficient on their own.

A central tenant of gentrification is its disproportionate impact on vulnerable communities and its contrast with historical neighborhood characteristics. Following World War II, mortgage guarantee policies passed in the GI bill helped subsidize Americans to move from metropolitan areas to suburbs. However, actions taken by the Federal Housing Administration rendered these suburbs largely unavailable to black Americans (“Urban Displacement Project”). The resulting migration flows, known as white flight, spurred a trend of disinvestment in inner city communities. It is only in these previously disinvested neighborhoods where gentrification may occur. The origin of this disinvestment also informs our identification of such communities.

Ding et. al (2016) define neighborhoods as *gentrifiable* if their median household income is below the citywide median at the beginning of the analysis time period and *gentrifying* if it was gentrifiable at the beginning of the period and experienced a percentage increase in median gross rent or median home value above the citywide rate and an above citywide increase in percent of residents with a college degree. A similar specification by Meltzer (2017) includes only average household incomes to identify such neighborhoods. These specifications identify many neighborhoods which are gentrifying, but fail to capture the extent to which gentrifiable neighborhoods have been deprived of social capital and investment. To better capture these housing dynamics, Freeman (2005) includes the share of the housing stock built in the last 20 years and restricts gentrifiable neighborhoods to only include those whose share of new housing was below the 40th percentile of their metropolitan area. Freeman also restricts their sample to neighborhoods within the central city of metropolitan areas where real housing prices increased over the period of analysis.

Bates (2015) adds a final layer to this analysis by incorporating social capital measures, including the a tract’s proportion of owner occupied housing units, and the percentage of residents who identify as non-White. In both cases, a tract must have a larger proportion than the citywide rate to be considered gentrifiable. This additional further detail limits the universe of gentrifiable tracts, but better encapsulates the historical context of urban gentrification. It is this, more detailed, view which our study aims to emulate.

# Literature Review: Crowdsourced Data

Forthcoming.

# Methodology

Following in the footsteps of many others (Meltzer 2017, Lester and Hartley 2014, Ellen and O’Regan 2008, McKinesh et al. 2010, Freeman 2008, Hwang 2014) our level of analysis will be the census tract. Census tracts contain approximately 4,000 residents and are frequently used to model neighborhood evolution. Demographic data is sourced from the American Community Survey (ACS) 5 year estimates via the Census API, which provide detailed information at the block group level.

As with the authors in sections two and three, our analysis of gentrification begins with identifying gentrifiable tracts. These will be defined as tracts with median household incomes below the county’s 50th%ile, educational attainment below the county’s 50th%ile, and median rent below the county’s 50th%ile. Our Airbnb data is already limited to the cores of metropolitan areas, so further geographic filtering by geography is not needed.

To measure gentrification, each gentrifiable tract will be given a score composed of an equally weighted average of the following variables, downloaded from the ACS: median household income, median monthly rent payment, percent of residents who identify as White and non-Hispanic, and percent of residents with a bachelors or higher degree. Changes in rent will be adjusted by the metropolitan statistical area’s consumer price index, as reported by the Federal Reserve. The percentage change a tract’s gentrification score at the outset of the study (2009) and its conclusion (2019) will be our dependent variable.

To explain tract variation, we propose two multiple linear regression models: one with structured Airbnb data and one with the unstructured. Independent variables in the former will include the tract’s population, number of Airbnb’s, the number of Airbnb Reviews, the Airbnb star rating and Airbnb location rating. We expect that a higher density of Airbnb’s will correlate with changes in neighborhood characteristics that make a tract more appealing to gentrifiers. The more appealing these characteristics are, the higher we expect the Airbnb location rating to be, thus we expect both the number of listings and the location rating to be positively correlated with gentrification.

In the second specification we include a matrix of unique words used in a given tract’s Airbnb listing reviews and their prevalence. We also hope to include a measure of review sentiment, although the feasibility of this piece is still being determined. These covariates will be added to the first model specification. We predict that the textual understanding of a neighborhood, as given by user reviews, contains useful information about the neighborhood demographics, and thus expect the matrix to increase the predictive power of our model. We do not have hypotheses at present about which sorts of words will have high explanatory power.

# Methodology Appendix

Additional documentation and detailed code to recreate this project is available at: <https://github.com/LeoKitchell/SeniorThesis>

# Results: Structured Data

Forthcoming.

# Results: Unstructured Data

Forthcoming.

# Implications and Policy Tools

Forthcoming.

# NOTES

Following in the footsteps of other empirical studies of gentrification, the unit of analysis will be census tract, using ACS 5 year data. A further examination using block groups may be conducted, time permitting, but we hypothesize that insufficient density of Airbnb's will rend the predictive power of the textual data minimal in all but the most dense tracts. To reflect the multi-stage nature of gentrification, our analysis will first identify tracts that are gentrifiable.

Like Bostic and Martin (2003), Freeman (2005), and McKinnesh et al. (2010 \*\*CHECK\*\*), getntrifiable tracks will be defined as those with a median income less than 50% of the median income of the metropolitan statistical area in which they are contained.

Based off of Jain (2021), but instead of reducing the dimensionality of the text data in LDA and Doc2Vec we keep the full variation of text by using a random forest regression methodlogy to test the significance od diffferent words in the review.

Leading scholars describe gentrification as a staged process (Hoover and Vernon 1959, Bates 2013, Goetz 2019) that

Although there is some debate surrounding how neatly these staged models can reflect

Suburbs have become less appealing to high wage workers as their leisure time has diminished. This has increased demand for urban housing stock, reversing a 1980s trend of suburbanization, and causing housing proximate to central business districts to balloon in price (Edlund 2015). At the same time, decades of "urban revitalization" projects are coming to fruition \*\*(citation needed)\*\*, further increasing metropolitan neighborhood home prices. Consumer tastes are also changing \*\*(citation needed)\*\* as new workers demand to be in culutural cities (potential research: are young tech workers making different housing choices than young, rich workers from the 50s-90s). Crime rates in cities have also fallen dramatically since the 90s, further bolstering home prices in urban areas \*\*(citation needed)\*\*.

### Different Specfification of Gentrification

Glaeser(2020) defines gentrifying areas as "places with high initial poverty, relative to the city as a whole, and high rent growth." They also use a secondary specification which inclues an "increasing education and increasing share of white residents" as measures of gentrification.

Jain (2021)

Slater(2006;2009), Newman and Wyly (2006), Davidson and Lees (2005) argue that displacement is "a constituent element of gentrification" (Goetz). In this view, displacement must be included in the gentrification specification to properly measure its effects. However, as Goetz et. al. write, "empiracally establishing the degree of displacement is one of the most difficult aspects of gentrification research...Even when data exist that might highlight one form of displamcement, typically most of then other forms of displacement remain largely hidden."

In this project I plan to use listing data on Airbnbs in major US cities to determine neighborhood features associated with gentrification. While these data may be able to pick up demographic changes in neighborhoods, I believe they will provide the most use in quantifying cultural change. This quantification will be a function of both structured features, such as the density of listings and their prices, and unstructured features, namely the text descriptions of Airbnb listings and their neighborhoods. \*If the markers of cultural change are indeed visible to residents, the text description of a given Airbnb and its neighborhood should reflect that change.\*