How are Businesses Prejudiced?
An In-depth Look at Commercial Microsegregation in Chicago

Adam Shelton

April 2019

1 Introduction

Segregation is perpetuated by a complex set of relationships, including those between living costs, wages, crime, schooling quality, and the desire to be with others of a similar culture to oneself [5]. Due to the multitude of factors behind segregation, and the complexity involved with measuring segregation in its own right, most studies of segregation pick a handful of different measures to analyze. Most often, studies link race to certain outcomes, like political affiliation and local amenities [7], or foreclosure rates [3] as it is much easier to do.

While segregation is easily noticed, it is much harder to quantify, as even a highly diverse neighborhood (or street) with an equal number of each race could still be segregated. Complicating the topic further, people not only segregate along racial lines, but also along age, income, and education levels [1]. Integrated neighborhoods are actually cheaper, with both blacks and whites having to pay more to live in a segregated area. This contradicts the theory that economics is the driving force behind segregation [1].

Only in recent years have scholars begun to study these smaller scales of segregation, called microsegregation or street-level segregation. In many cases, while a neighborhood is diverse, the places that inhabitants of that neighborhood visit remain segregated [4]. People of different backgrounds, while residing in the same neighborhood visit different social institutions and businesses. I aim to build on this corpus by studying how businesses contribute to microsegregation in Chicago.

2 Data

Traditionally, demographic data has been used to research segregation. Demographic data, while collected quite meticulously and with comprehensive granular data of large areas, has also only covered very simple measures of these areas. This has been slow to change, as much of the demographic data is collected by governments in censuses. Although the US Census as of late has begun to measure a much wider range of topics, including those relating to internet access and technology in the American Community Survey of the past few years [2]. Where segregation falls along neighborhood or census tract lines, demographic data can perhaps capture segregation well. However, this is unlikely to be the case, as even diverse neighborhoods can be segregated by race or income, creating low street diversity and segregated businesses and amenities [4].

However, I propose that digital mapping data from Open Street Maps offers a much more comprehensive coverage of geospatial dimensionality. Modern digital map platforms cover a highly granular set of data, giving names and categories to places down to individual buildings. This not only captures amenities and businesses, but especially in urban areas, residential complexes as well. This allows models to capture not only the types of buildings but also specifically named buildings, such as businesses with several branches or chain stores. These mapping services also display roads, waterways, train tracks, and other landmarks that oftentimes create physical boundaries between segregated neighborhoods.

Open Street Maps also offers historical data, which stretches back to 2012 for the entire planet, creating a data-set approximately one terabyte in size [6]. Digital mapping services contain a virtual treasure-trove of data, especially for this purpose, as the types of buildings in any given area are shaped by a myriad of cultural, economic, and regulatory forces, which are difficult to capture on their own.

While I plan to limit my scope to the Chicago metropolitan area, as digital mapping services cover large amounts of the developed world, this data can be used to measure segregation in any region across the US, or feasibly, the entire world. This would allow more accurate comparisons of segregation, which can allow further enlightenment on how specific policies may effect street-level segregation, which would be difficult and time consuming to do by more traditional methods.

As Open Street Maps data encompasses the entire planet [6], it must be subset to make it usable, especially for this application. I initially chose a 73 by 102 km (45 by 63 mile) rectangle from 41.3954°N, 88.1117°W to 42.3129°N, 87.2411°W, an 1.8 million acre section of northwestern Illinois. This effectively covers the Chicago metropolitan area, although the exact area can be altered if necessary. This chosen area encompasses 946,247 "multi-polygon" locations, of which 886,871 are defined as being buildings. 25,544 of these locations are named, and 20,636 have one of 107 unique amenity categories attached, such as bank, hospital, cafe, fast food, etc. Other tags for each of these locations includes address information, if the place has a building, the type of sports facilities available, the types of offices or shopping available, if the place may be of interest to tourists, etc.

While the data is large and comprehensive, it is also quite messy. For example, according to the data, there are only three ATM's in the dataset according to the ATM amenity category, which is almost certainly incorrect. Cleaning and verifying almost one million data points is sure to be a significant challenge, if not virtually impossible, which may present an issue with the accuracy of the models generated from it.

For more about the subset of Open Street Map data used, see the 'Data Visualizations' section.

3 Analysis Methodology

Digital mapping data is large and highly multidimensional, which necessitates a different strategy for analysis. As such, I am proposing that this mapping data is analyzed in a non-traditional manner, using unsupervised learning methods. Traditional modeling methods used in the social sciences, such as linear regression, impose a specific form on the relationship between the independent and dependent variables. In the case of a linear regression model, the model imposes the form of a linear relationship. However, this relationship may not be correct to assume, and in the case of the complex co-linearity and interactions that make up many previous segregation and gentrification models, it is probably correct to assume that these relationships are in fact highly non-linear. As unsupervised learning methods assume very little to no functional form in the modeling process, it makes them uniquely suited for the task of modeling segregation and gentrification.

Furthermore, as segregation is simply a grouping of people with similar traits, unsupervised clustering methods would likely have very good performance in modeling segregation, while still capturing the high dimensionality associated with it. Since they assume no functional form or particular outcome, they will capture the actual level and relationship of segregation, instead of traditional studies that only study segregation as it pertains to race or socioeconomic class. This also reduces the bias of the researcher in the analysis, as the model shows the relationship in the data, despite what ideas the researcher is predisposed to.

While, unsupervised clustering does reduce interpretability, I believe that enough explanatory information can be gleaned from feature importance and surrogate models. This would undoubtedly result in a less precise answer to the degree to which specific businesses or categories of business are segregated in Chicago, but it would still likely allowed for an accurate comparison among them. This would result in a much more insightful analysis, as the analysis does not need to be restricted to a small area or a specific selection of businesses.

References

[1] David M. Brasington, Diane Hite, and Andres Jauregui. House price impacts of racial, income, education, and age neighborhood segregation. *Journal of Regional Science*, 55(3):442–467, 2015.

[2] US Census Bureau. Computer and internet use. Accessed: 2019-04-09.

[3] Matthew Hall, Kyle Crowder, and Amy Spring. Neighborhood foreclosures, racial/ethnic transitions, and residential segregation. *American Sociological Review*, 80(3):526–549, 2015.

[4] Derek Hyra. Greasing the wheels of social integration: Housing and beyond in mixed-income, mixed-race neighborhoods. *Housing Policy Debate*, 25(4):785–788, 2015.

[5] Trevor Kollmann, Simone Marsiglio, and Sandy Suardi. Racial segregation in the united states since the great depression: A dynamic segregation approach. *Journal of Housing Economics*, 40:95–116, 2018.

[6] Open Street Maps. Planet osm. Accessed: 2019-04-09.

[7] Jessica Trounstine. Segregation and inequality in public goods. *American Journal of Political Science*, 60(3):709–725, 2016.

4 Data Visualizations

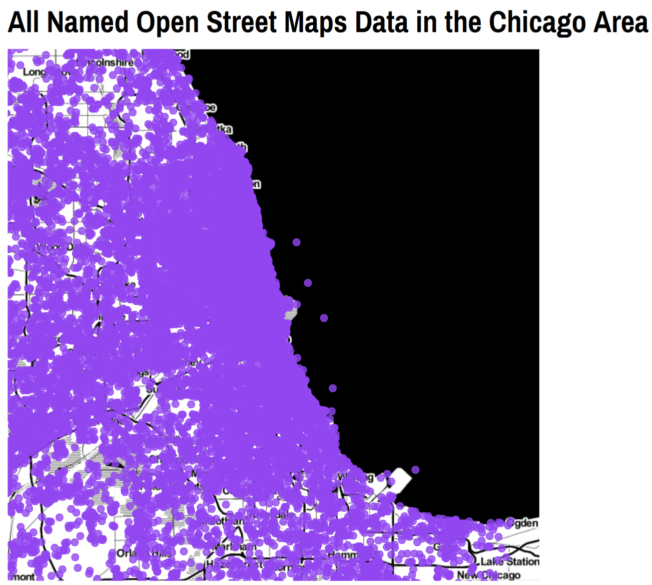

Figure 1: Even just the 25,000 locations with names cover the entire city

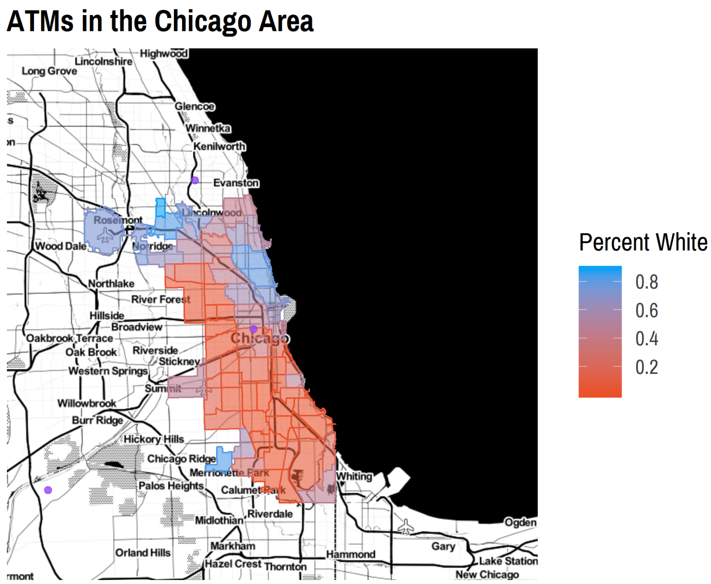

Figure 2: The data does not appear to have ATM's coded correctly

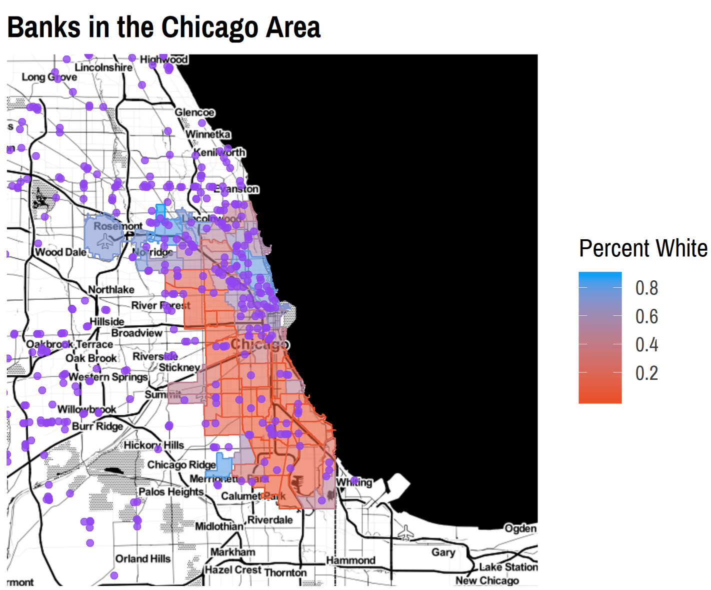

Figure 3: More banks appear to exist in whiter areas

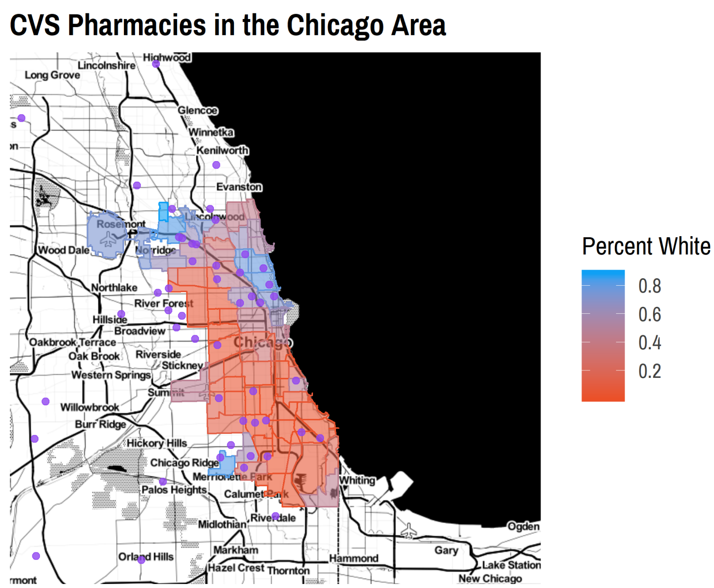

Figure 4: CVS pharmacies do not cover the entire city equally

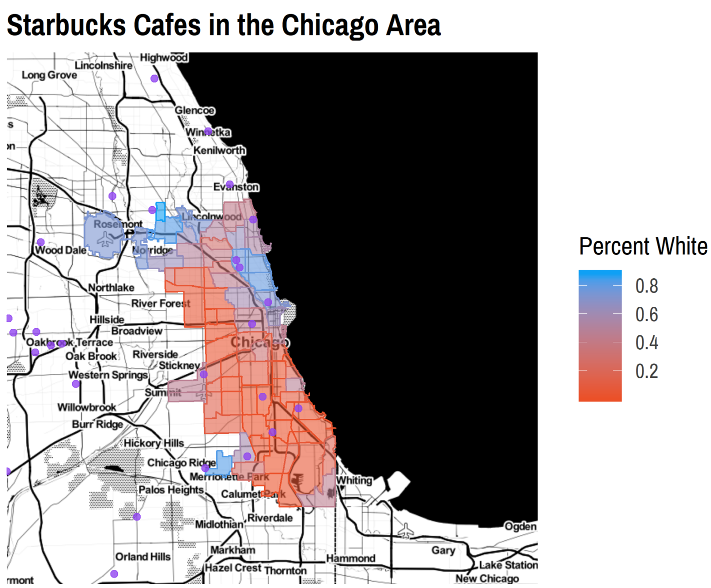

Figure 5: Starbucks are only located in very specific areas, but not necessarily in a racialized way