Commercial Segregation in Chicago

Adam Shelton

June 12, 2019

1 Background

Chicago is notorious for its segregation, a result of many years of racial injustices, which continues to perpetuate inequality in the city to this day. This rift in the city between the white communities of the North Side and the communities of color in the South Side has made a city of two worlds, with one having considerably more poverty, violence, and opportunities than the other. Despite the end of de jure segregation in the twentieth century, these residential clusterings have changed very little, even with the numerous efforts aimed at doing just that.

While segregation is primarily thought of as being a racialized residential phenomenon, there is evidence that the problem is not quite as simple as people living with other individuals of a similar skin color to themselves. Segregation is not just exclusive to race, but overall people divide themselves on numerous divisions, whichever will get them the most utility [Brasington et al., 2015]. Separated communities are 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 [Kollmann et al., 2018].

The reasons for segregating are not always financial either, as typically living in a segregated area is more expensive than living in an integrated community [Brasington et al., 2015]. 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. Studies have linked link racial segregation to many outcomes, including political affiliation and local amenities [Trounstine, 2016], or foreclosure rates [Hall et al., 2015]. This has serious ramifications, leading to segregated social networks, less social mobility, and awareness of the needs of others [Tach, 2014]. Lack of interaction with others due to Segregation can also reinforce stereotypes, which are used by outsiders to inform the opinions their opinions of segregated areas [Hwang and Sampson, 2014].

These issues are compounded by inequality in businesses in segregated communities. Black business owners tend to have less capital, less education, and less access to loans, resulting in smaller, less successful businesses on average [Bates, 1989]. Entrepreneurs of color also have less social capital, in part due to the significantly smaller proportion of black business owners to other ethnicities [Wingfield and Taylor, 2016]. Undocumented workers face punitive challenges in starting their own businesses, and all minority workers face systematic discrimination throughout the life of their business [Romero and Valdez, 2016]. Immigrants on visas waiting to become permanent US citizens often turn to self-employment, not just to help pay the bills, but also to reconnect with the independence they had before they moved to the US [Banerjee, 2019].

Yet, despite while Chicago has a high proportion black businesses, only topped by New York City in the United States, black business owners in Chicago faced similar issues as elsewhere [Harper-Anderson, 2017]. Many of the entrepreneurs mentioned struggles with obtaining capital, maintaining strong business relationships, and even their motivation for starting their own businesses [Harper-Anderson, 2017]. Throughout Chicago, and America as a whole, segregation and discrimination affects minorities not only in their homes but their businesses as well. This project aims to investigate these claims further to identify if businesses are segregated in similar spatial patterns as residents, and what might be contributing to these separations. While our findings warrant further research on the matter, it appears that businesses are segregated in meaningful ways which are tied to cycles of poverty (or lack there of), leading to continual investment disparities among segregated communities.

2 Traditional Problems

Much of the past research withing these realms have been problematic for a myriad of reasons, of which this project attempts to address head-on. The most glaring of these issues is that traditionally segregation has been considered purely racial in nature, or at best by socioeconomic class. With very little reasoning, it is quite obvious that many factors contribute to segregation. Age segregation is exhibited at small scales in gated communities for senior citizens, or the tendency for younger people to gather in urban centers with older people calling the suburbs home. Families with the means to do so move to communities with the best schools, childcare, and safety. While there is certainly nothing wrong with selecting one of these facets of segregation to study, the highly interactive nature of these different elements allows for only a very narrow scope of analysis. If our goal is to study these segregational structures as a whole, we must account for every aspect of it.

Additionally, until very recently, most segregation research has failed to account for geographic scale. A neighborhood that appears to be integrated, may really be segregated on a smaller scale, also known as micro-segregation. Most segregation studies have traditionally been conducted using Census data at the tract level [Lee et al., 2008]. The popularity of the census tract can be "traced in part to its convenience and to the belief that it approximates a 'real' neighborhood" [Lee et al., 2008]. However, different patterns of segregation are visible at different scales, which affects not only measures of segregation, but how outcomes of segregation are studied [Reardon et al., 2008].

Lastly, the vast majority of segregation research has focused on the individuals that make up a community. Yet, businesses are an equally important reflection of any community, providing services and jobs for the residents who live there. Even in diverse neighborhoods, social groups stay in their own section of the neighborhood, only patronizing nearby businesses, with their children attending separate schools [Tach, 2014]. Only a few parks, and most convenience stores, attracted a diverse clientele [Tach, 2014].

3 Data

Two data sources are used for this analysis. 2010 Decennial Census data is used to generate a proxy for segregation. While past research has primarily focused on racial segregation, there exists many dimensions for which people are segregated, including race, socioeconomic background, type of family, profession, etc. Census Data can be used to capture many of these dimensions.

Open Street Maps is an open source alternative to modern mapping applications like Google Maps, providing information on the features that define our physical world, including roads, boundaries, and places. This project specifically uses place of interest data from Open Street Maps to describe businesses and their locations within Chicago.

3.1 Census Data

2010 Decennial Census data was retrieved through the public Census API for the entirety of Cook County, Illinois using the TidyCensus R package. The data includes not just census variables, but also shape data defining the geometries of each region retrieved. For this project, data from 1319 census tracts in Cook County, Illinois were used. While, 39 of variables were downloaded using the Census API, the 24 variables used in the analysis are noted in the table below. This excludes two variables identifying each census tract, and one variable containing the tract geometry, used for the geospatial analysis. Any raw population counts, like the number of people identifying as white in a tract, were converted to percentages by dividing by the total number of people in the tract. Variables that were already measured per capita, like median age or average household size were left as-is.

It is difficult to describe geospatial data purely by the numbers, so visualizations were generated using the ‘ggplot2’, ‘sf’, and ‘ggmap’ packages in R. Chicago is notorious for its racial segregation, stemming from racialized efforts, such as redlining, in the twentieth century. Despite its overall racial diversity, very little of the city lives in integrated communities, exhibited in census data has the high racial homogeneity in tracts. Tight, defined pockets of each race are clearly visible, with whites primarily residing on the North Side of Chicago and areas outside the city, and blacks and

Latinos having their own respective areas on the South and West Sides, with very little blending in-between.

Other notable relationships included the proportion of rented units, and the percentage of minors in a tract. High proportions of rented units are found in a uniform radius immediately surrounding around The Loop. High proportions of minors are primarily found in North Side tracts, displaying a similar pattern as proportions of whites.

The proportion of minors is negatively correlated with the proportion of one person households, which makes sense, as typically minors cannot live on their own. Median age is highly correlated with the proportion of owned housing units, indicating that older people are more likely to own a home.

The proportion of households with both a husband and wife is positively correlated with proportions of home ownership and whites, while being negatively correlated with proportions of rentals, vacant houses, and blacks. Tracts with higher proportions of blacks accompanies a significant rise in proportions of females and vacant houses, which speaks to the socioeconomic disparities between races, as well as perhaps the greatly increased incarceration rates of men in communities of color.

The proportion of Latinos in a community has a strong relationship with the percentage of people in that community who identified as Native American or two or more races, and a perfect correlation with those who identified their race as other. This shows the confusion surrounding the separation between race and ethnicity, especially as it pertains to Latinos, who often feel they do not fit into historically defined race categories.

Overall, 2010 Decennial Census data provides adequate demographic data to capture different types of segregation in Chicago. While the Decennial Census does not include data on dimensions such as income or employment, which certainly play a role in segregation, it does provide a much higher degree of accuracy and completeness than other demographic surveys such as the American Community Survey. In addition, while there does not exist a single variable for certain dimensions, like income, it is likely to be captured to some degree through a combination of other dimensions, like race and housing status.

3.2 Open Street Maps Data

Open Street Maps data encompasses the entire planet, making it possible for use in geospatial research anywhere, but its size necessitates must be subsetting to make it easier to work with. A 73 by 102 km (45 by 63 mile) rectangle from 41.3954°N, 88.1117°W to 42.3129°N, 87.2411°W was chosen, covering a 1.8 million acre section of northwestern Illinois. This effectively covers the majority of places within Cook County. This chosen area encompasses 946,247 multipolygon 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. Another 111,516 locations were stored as points, resulting in over one million locations available in the raw data. 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. This data was downloaded through the export tool provided by the organization at `https://www.openstreetmap.org/export`.

As this data is downloaded in an open but proprietary format, there were some unique challenges in importing the data into R for analysis. Packages like 'osmdata' have been developed to not only give R the ability to read the proprietary OSM files, but to also query the Open Street Maps API directly, allowing researchers to forgo the process of manually downloading the files from the Open Street Maps export tool. Unfortunately, neither method with the package appeared to work, perhaps because the data-set was too large for the third-party package to handle. Therefore, 'ogr2ogr' an external tool as a part of the open-source GDAL geospatial translation library was used to convert the OSM file into a standard shapefile, which could be easily read using the 'sf' package in R.

As part of the conversion process done by 'ogr2ogr' the data was split into separate shapefiles for each type of geometry in the original OSM file (lines, multi-line strings, multi-polygons, and points). This necessitated the joining of the points and multi-polygon data, which would encompass all buildings, and therefore businesses. Another effect of the conversion process meant that many of the variables in the dataset were stored as a vector of key and value pairs converted to a characters.

Using R, a selection of these variables with the least amount of missing values were parsed out, to get the variables necessary to conduct the analysis.

While both the points and multi-polygon data-sets are organized slightly differently, and both have many missing values, all Starbucks locations in the two data-sets were analyzed to determine whether the data-sets were likely to have repeated entries among them. Address data was used when available, and coordinates compared when not, with locations verified by comparing any suspected duplicates to Google Maps as an extra layer of verification. Starbucks were chosen as they are one of the most common businesses in Chicago, are spread across the city, and were thought to be less likely to be in close proximity to each other, avoiding the likelihood of falsely categorizing two close locations as the same location. Unfortunately, due to the missingness in the data-sets, and no common unique identifier between the two, it is virtually impossible to determine with absolute certainty that the their are no duplicate values. However, under the assumption that due to the crowd-sourced nature of the data, and no evidence of any duplicate Starbucks locations, these two data-sets were merged together as if they were completely unique.

This merged data-set was then filtered by name and amenity to get 10,359 businesses across the Chicago area. Any observations without data for these two variables was discarded, as the theoretical framework suggested that certain businesses and types of businesses are likely to be segregated. Locations were filtered by amenity, removing buildings such as schools, police stations, and parks, which would not be considered businesses. Ideally, other variables, including those for the street and zip code of where the business was located, or if the business was a shop or retail would have been included, but filtering for observations with those variables would have reduced the size of the data-set even further. This processing results in 4 variables being included in the final data-set used for the analysis: latitude, longitude, name, and amenity, with values for every observation.

4 Analysis Methodology

This project uses clustering models for the bulk of the analysis, as segregation by definition is a clustering of alike people in one geographic area. Unsupervised clustering models are used where possible to give an unbiased description of segregation in a given area. This allows for the model to form relationships out of the data, rather than the traditional approach to studying segregation, where a specific functional form and segregation outcome (usually racial segregation) is imposed. However, as exhibited in a cursory visual analysis of census data, even in Chicago, a city famous for its racial segregation, other types of segregation exist as well. This necessitates a broader definition of segregation, and therefore, an unsupervised modeling method, to improve accuracy in capturing the immense multi-dimensionality exhibited by segregation.

The cost of this increased accuracy of unsupervised learning methods, is a trade-off in interpretability. Unsupervised clustering methods are black box models that give little insight into the factors contributing to the model. This is rectified through the use of a cross-validated decision tree surrogate model. These model analysis methods were chosen for their ease of computation and understanding.

4.1 Demographic Clustering Models

All variables except for the two identification variables, and the tract geometry are used in the model. This means that no spatial data such as coordinates or geometries are used in the training of the model, just demographic data. The data is centered and scaled across the data-set prior to training, and any observations with missing values dropped.

Two unsupervised clustering models are used with the data, K-means clustering and Hierarchical clustering. Since there are no ways to directly validate the accuracy of the models, all the data is used for each model. Multiple models are trained, each at a different number of clusters, between two and six, and the outcomes visualized. Cluster sizes from one to ten are used when generating fit estimates across different cluster sizes to help determine the best cluster size to use. Both these models were trained using their respective functions in the R 'stats' package.

In addition to visually comparing the output of the unsupervised clustering models to the theoretical expectations, measures of within cluster sum of squares, silhouette width, and gap statistic were used to help determine the "optimal" number of clusters for each model. All of these analysis of the possible model outcomes were used to select a five cluster model as the optimal model, due to its performance on the aforementioned tests, and a balance of specificity to interpretability. Too many clusters made it more difficult to determine what was driving the model, while two few clusters gave too little information. The number of clusters was kept small as Chicago is highly macro-segregated [Lee et al., 2008], and macro-segregation in Chicago accounts for more than 70% of the micro-segregation [Reardon et al., 2008], meaning there is more likely to be a few large clusters rather than many smaller clusters.

The outcomes from the final model are then joined to the un-scaled data and used to generate a surrogate model. A cross-validated decision tree was chosen for this model, as it is an modeling method for a categorical outcome that is both easy to compute and easy to understand. Decision trees are highly variable, however, as this it is being used as a surrogate model to help understand which factors are likely contributing to the model, the general themes inferred from the model can still be gleaned, even if the specific features and their weights vary between models. The decision tree model was built using the “rpart” model from the ‘caret’ package in R. The modeling process takes advantage of ten-fold cross validation repeated three times, with an automatic grid search of length 10 used for hyperparameter tuning.

4.2 Business Clustering Models

To study the segregation of businesses in Chicago using Open Street Maps data, the process is very much the same except for a few key differences. Both K-means and Hierarchical clustering models are used in the same manner as with the demographic data, as is the analysis method a decision tree surrogate model. However, as the amenity and name data consists exclusively of text data, those were first converted to numeric factors before the data was scaled. In addition, while the demographic models used no location data, by necessity, point coordinates must be included in the business model. The types of businesses and even many specific businesses are not unique to any one area, but the number and proximity of those businesses likely is.

4.3 Visual Analysis

All the visualizations were made using R, primarily using the 'ggplot2' package. The choropleths of Chicago used the 'ggmap' package to generate background black and white map tiles from Stamen (`https://stamen.com/`). Cluster principal component plots, and optimal cluster size plots were made using the 'factoextra' package. Correlation plots used the 'ggcorrplot' package, and decision tree plots used the 'rpart.plot' package.

5 Results

Overall, the models produced mixed but still promising results, allowing the research question to be partially answered. It appears that businesses are segregated in similar ways as residents, but the reasoning behind it is unclear. Still, the bulk of these inconsistencies appear to be the result unexpected technical failures, not necessarily issues with the methodology.

5.1 Demographic Cluster Model

The two types of clustering models resulted in drastically different outcomes for the demographic data. The K-means model appeared to cluster the data quite well at each cluster size. At two clusters, the groupings seem to most resemble race, with a central cluster covering most of the South and West Sides, and a larger cluster covering the North Side and areas surrounding the city. The two clusters are also well separated for the first two principal components. At three clusters the result is similar, except now it appears to be split not just by white/non-white but white, black, and Latino.

Naturally, as more clusters are added they separate less clearly on the first two principal components, but at four clusters, Chicago's 'sides' are pretty clearly separated from areas surrounding the city, with five clusters separating more of the surrounding areas. Until five clusters, the K-means model appears to be highly reflective of race, indicative of the strong racial segregation in the city.

Interestingly, tracts making up the Hyde Park and Kenwood neighborhoods are separated from other communities on the South Side at every cluster size. The model consistently groups Hyde Park and Kenwood with tracts on the North Side. This makes a lot of sense, for while Hyde Park and Kenwood are located on the South Side, they stand apart from other nearby neighborhoods, being much more racially and socioeconomically diverse, and gentrified, in large part due to the University of Chicago being located there. Colloquially, this area is often referred to an "oasis" or "island" within the South Side, due to the significantly lower rates of poverty, crime, and violence compared to surrounding neighborhoods, and the fact that this is reflected in the model gives some strong evidence for its validity.

Unfortunately, the hierarchical clustering model under the default options did not garner the same success, with the virtual entirety of Chicago belonging to the same cluster. However, changing the model to instead use the squared Ward algorithm ("ward.D2" with the 'hclust' function) appears to garner the expected result, and is quite similar to the K-means model. However, the overlap between clusters on the first two principal components appears greater than that of the K-means model, suggesting the K-means model is doing a better job of separating the clusters, even if the outcome is relatively equivalent.

Both decision tree surrogate models select many of the same variables for its decisions. As a single decision tree can vary greatly between different builds of the same independent variables, there is little value to splitting hairs on the specific differences and values, as it will change slightly each time. However, both surrogate models indicate that, unsurprisingly, race is a significant factor in these models of segregation, more interestingly, variables such as age, the percent of households with two parents, percent of families, and housing status all contribute to the model. In the hierarchical surrogate model, the decision tree indicates that the primary differentiation between black communities in on the South Side, and North Side communities is just the proportion of black people. Otherwise, these two communities can be quite similar, having a similar proportion of families, whites, Latinos, and average family size.

5.2 Business Cluster Model

Both the K-means and Hierarchical clustering models had a much harder time separating the data, despite the fact that coordinates were included as independent variables. The K-means model did not begin to cluster geospatially until a cluster size of three, and even on models with more clusters, there was considerable overlap of the clusters with the first two principal components. At five clusters, the K-means model separated the far Southwest and Northwest sides as two distinct clusters, but the other three clusters were a jumbled mess in the center of the city. Adding a sixth cluster only added to this central grouping of several different classifications. However, the density of businesses does seem much less on the South Side of this grouping compared to the near West and North Sides, but it is unclear if this is proportional to the change in population density between these areas or not.

The Hierarchical clustering model appeared to perform much better in comparison. At two clusters, the split is roughly along the center of the city, creating a Northern and Southern cluster. At five clusters, the South Side and Southwest sides have their own clusters for the most part, with the West side differing slightly from Southwest and Northwest sections of the city, and two categorizations blending together on the North Side. While this is a significant improvement over the K-means model, it does not offer the resolution that the comparatively less granular demographic data at the same amount of clusters, and only offers a passing resemblance to the residential segregation of Chicago.

Decision tree surrogate models were also developed for the business clustering models, however technical difficulties kept them from completing. Perhaps due to the increase in observations over the smaller demographic data, or some bug in the code, R would run out of system resources and crash before the models could complete.

6 Conclusion

While the results are not irrefutable evidence for or against the hypothesis, models developed throughout this project do tentatively support the hypothesis. The best clustering model of Chicago businesses, which appears to be the hierarchical cluster model, does generally reflect the segregation of residents in the demographic models. The South and West Sides in both are distinct from the rest of the city, with the center of Chicago appearing to be an intersection of many different clusters, and the suburbs north of the city are far less homogeneous. The demographic models capture this very clearly, while the business models are like looking at the same plot without glasses. Clearly the trends seem similar, if they are note exactly the same.

However, likely this is a problem with the data source rather than the modeling methodology itself. Open Street Maps data is highly multi-dimensional, as the goal of any digital mapping platform is to account for everything and anything in the physical world, delivering the most accurate and precise navigation results. Yet, while Open Street Maps is very good at telling us that something exists at a given spot, it is much less proficient at determining the qualities and information of that place, as exemplified by the large amount of missingness for aspects as simple as an address.

Part of this could also be the open source nature of Open Street Maps. While anyone can contribute to the platform and make it better, that does not necessarily mean that they will. Incentives are often needed when crowd-sourcing data to entice people to contribute more and better information. This is likely compounded by the big competitors to Open Street Maps, such as Google Maps and Apple Maps, that are far more popular among the general population. Not only do these platforms come installed by default on mobile platforms developed by each company, Google and Apple are some of the largest and most profitable companies in the world. They have much more money and resources to throw at not only improving their mapping platforms on their own, but advertising and giving perks to users for crowd-sourcing data for them.

Regardless of the reasoning behind the lack of data in Open Street Maps, the amount of missing data likely contributed to the difficulty clustering methods had in separating the data. There is only so much that can be done for two categorical variables plus two variables for location, especially when those categorical variables have hundreds of categories, as the ammenity and name variables had. Machine learning models, like these clustering methods, thrive on large, highly multidimensional data, as evidence by the ease with which they could pick apart the demographic data. A better data-set that is entirely comprised of businesses, perhaps like Yelp or business licences for a city, and that provide much more data about each individual business would likely perform much better, allowing for more defined clusters, and a better comparison to residential segregation.

Yet, despite these shortcomings, many parts of the project were successful, providing insight into segregation and inequality in Chicago. This novel use of machine learning does seem to be well suited for modeling segregation, offering a valid alternative to traditional methods. While all these methods are, to some degree, simply measuring the distance between data points, unsupervised clustering methods offer much greater ease of use, better accommodation of highly multidimensional data, and a higher chance to discover new relationships in the data, that would have been easily missed with a more strict functional form. Clearly, unsupervised clustering methods are not the be all end all of segregation research, but they are proving themselves as a valuable tool to use for such a venture.

However, even with the challenges faced throughout this project, there is still much to be learned from its results. It is clear from the demographic models that race is a significant component of segregation in Chicago, but not the only one. People who can choose where they want to live are situating themselves among others like themselves. Whether these are families with kids wanting to live in quiet areas with good schools, or single young professionals wanting to situate themselves in the heart of the city with invigorating night-life or simply whether people want to rent or buy their home.

Many of these different aspects of communities are also highly colinear, indicating that a few social forces are likely responsible for segregation than many. The proportion of blacks and whites is highly negatively correlated, a result of the racialized aspects of segregation. People in black communities are more likely to be renting not owning, with higher rates of vacancy, and slightly more young people. Families in these communities are more likely to be single parent households, and proportions of females higher. This is indicative of the higher rates of poverty, urban blight, and incarceration among these communities, providing evidence on why businesses there might be different than elsewhere in the city.

While the lack of a working surrogate model for the business models does restrict the amount of analysis possible, there is some evidence of inequality among commercial establishments in Chicago. Visually analyzing the six most common types of businesses in Chicago, it becomes apparent that while most of the city has a diverse range of business types, the South Side stands out for the opposite. The vast majority of businesses on the South Side are not banks, bars, or cafes, but gas stations and fast food. This presents serious implications in this region for two reasons. These types of businesses do not attract people from outside of the community to come spend their money in that community, reducing investment and jobs in the region. Secondly, these businesses do not provide decent paying jobs with good benefits, forcing residents from here to perhaps have to travel further for work than other areas of the city.

Additionally, if the proportion of fast food restaurants to grocery stores or other restaurants is higher in the South Side than other parts of the city, this perhaps has an effect on public health. Accessibility to affordable and healthy food is not only the basis for a long life, but can reduce health-care costs, which are already excessive, especially to those living pay-check to pay-check.

These issues are likely just the tip of the iceberg, with many more hiding underneath to discover. While this project has not been able to answer every question about the segregation of business in Chicago, it has offered evidence that it exists, and that there are distressing consequences as a result. There has been significant research conducted on the segregation of residents, and this project offers compelling evidence that more needs to be investigated about commercial segregation, and how it might be mitigated.

References

[Banerjee, 2019] Banerjee, P. (2019). Subversive self-employment: Intersectionality and self-employment among dependent visas holders in the united states. *American Behavioral Scientist*, 63(2):186–207.

[Bates, 1989] Bates, T. (1989). The changing nature of minority business: A comparative analysis of asian, nonminority, and black-owned businesses. *The Review of Black Political Economy*, 18(2):25–42.

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

[Hall et al., 2015] Hall, M., Crowder, K., and Spring, A. (2015). Neighborhood foreclosures, racial/ethnic transitions, and residential segregation. *American Sociological Review*, 80(3):526–549.

[Harper-Anderson, 2017] Harper-Anderson, E. (2017). Contemporary black entrepreneurship in the professional service sector of chicago: Intersections of race, entrepreneurship, and economic transformation. *Urban Affairs Review*, 55(3):800–831.

[Hwang and Sampson, 2014] Hwang, J. and Sampson, R. J. (2014). Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods. *American Sociological Review*, 79(4):726–751.

[Kollmann et al., 2018] Kollmann, T., Marsiglio, S., and Suardi, S. (2018). Racial segregation in the united states since the great depression: A dynamic segregation approach. *Journal of Housing Economics*, 40:95–116.

[Lee et al., 2008] Lee, B. A., Reardon, S. F., Matthews, S. A., O'Sullivan, D., Firebauch, G., and Farrell, C. R. (2008). Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73:786–791.

[Reardon et al., 2008] Reardon, S. F., Matthews, S. A., O'Sullivan, D., Lee, B. A., Firebauch, G., Farrell, C. R., and Bischoff, K. (2008). The geographic scale of metropolitan racial segregation. *Demography*, 45(3):489–514.

[Romero and Valdez, 2016] Romero, M. and Valdez, Z. (2016). Introduction to the special issue: intersectionality and entrepreneurship. *Ethnic and Racial Studies*, 39(9):1553–1565.

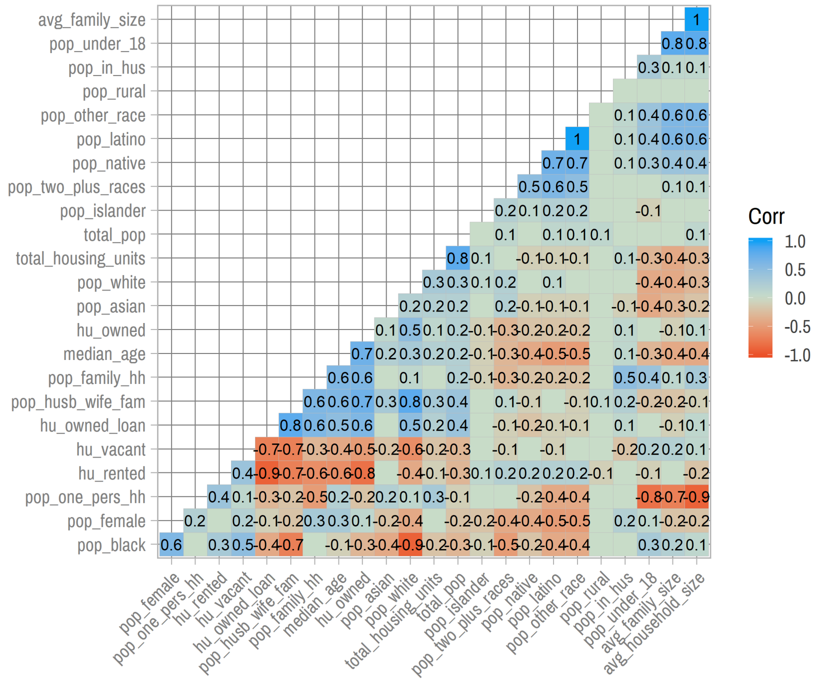
[Tach, 2014] Tach, L. M. (2014). Diversity, inequality, and microsegregation: Dynamics of inclusion and exclusion in a racially and economically diverse community. *Cityscape*, 16(3):13–46.

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

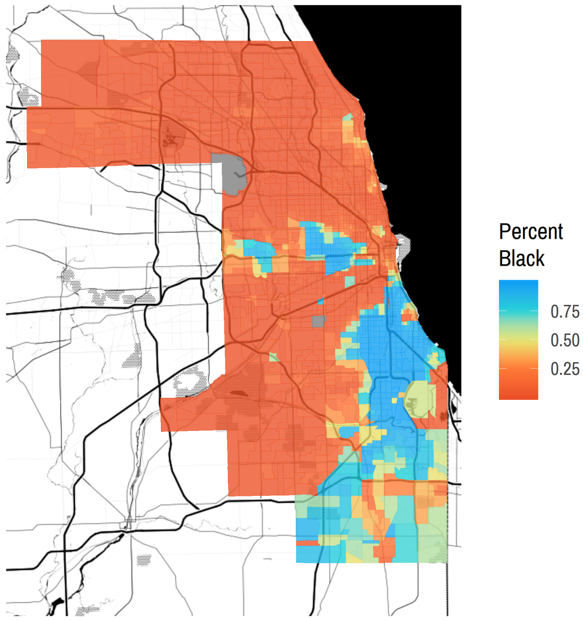
[Wingfield and Taylor, 2016] Wingfield, A. H. and Taylor, T. (2016). Race, gender, and class in entrepreneurship: intersectional counterframes and black business owners. *Ethnic and Racial Studies*, 39(9):1676–1696.

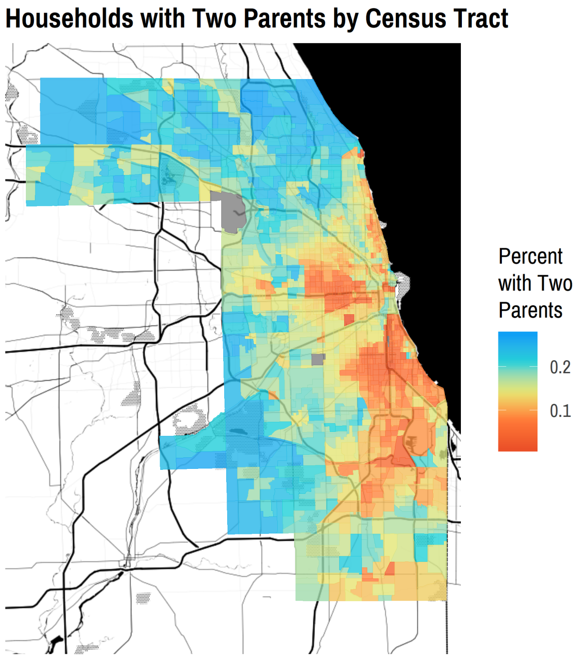
7 Tables and Figures

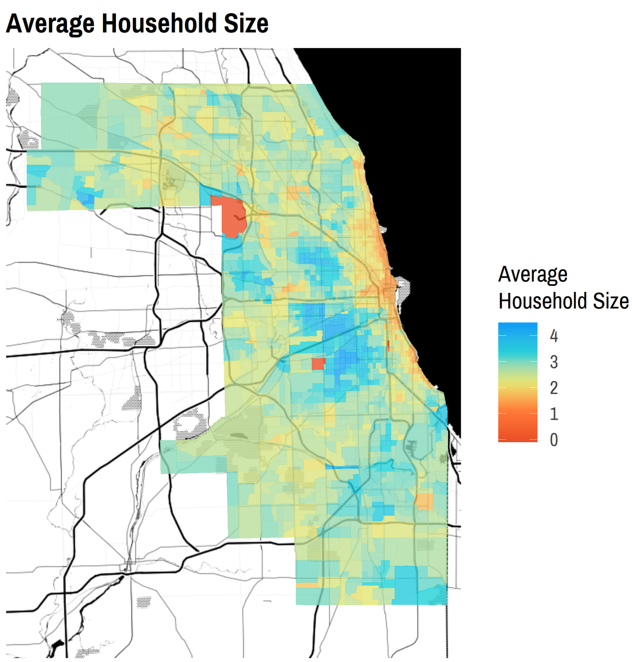
7.1 Demographic Data

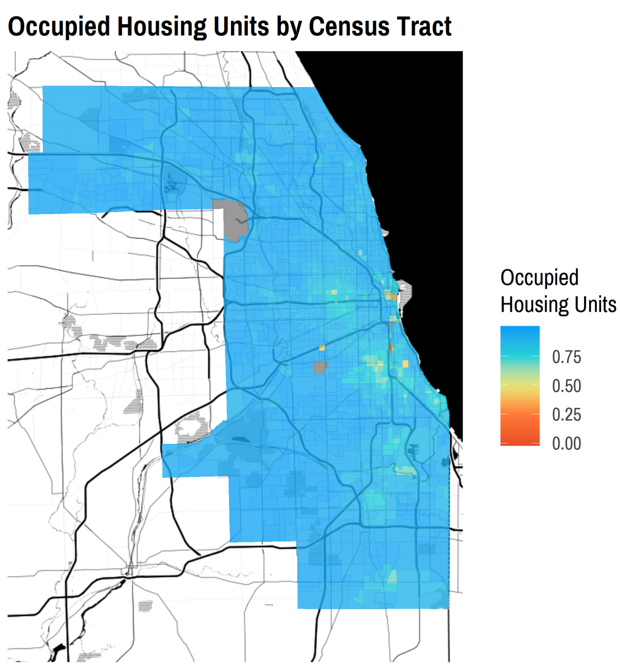


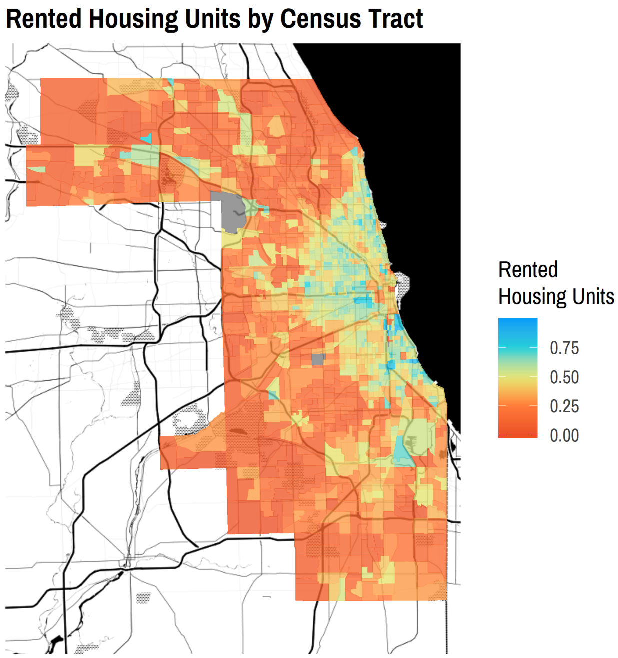
Percent Black by Census Tract

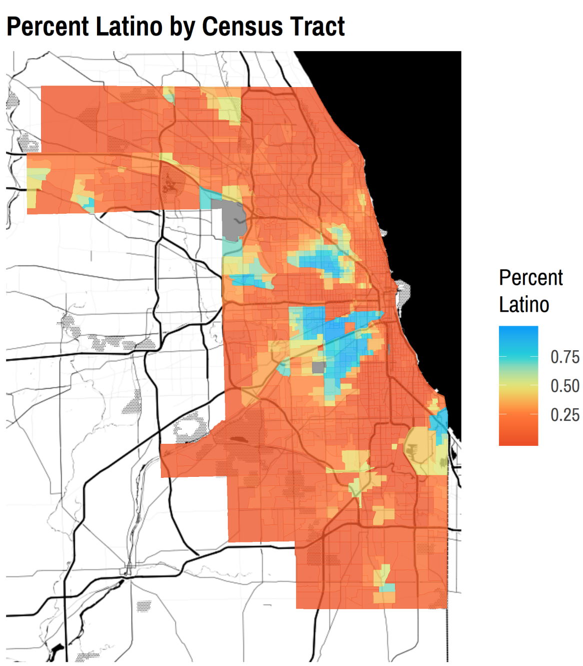


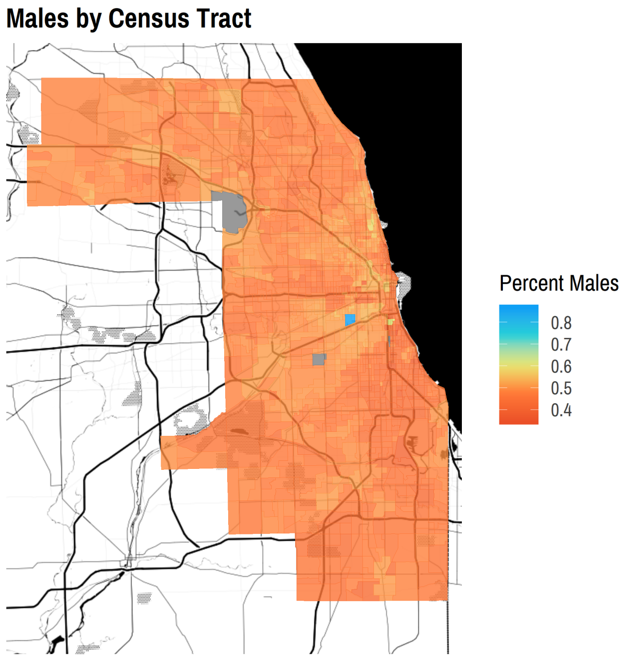

Households with Two Parents by Census Tract
Percent
with Two
Parents
0.2
0.1

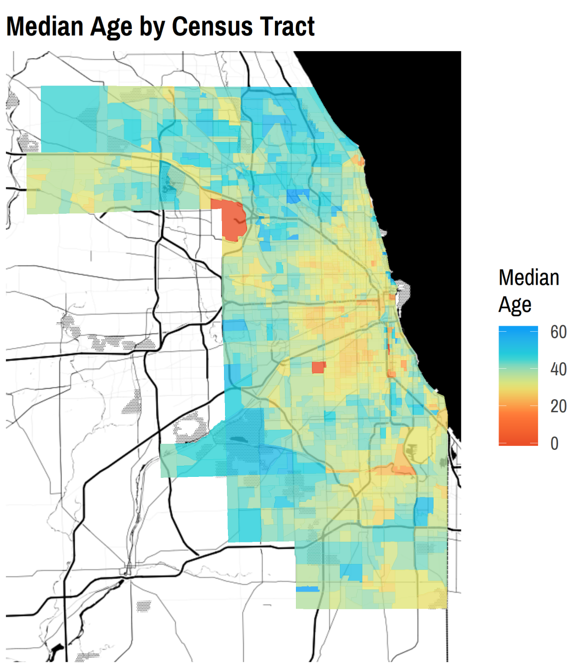

Average Household Size
Average
Household Size
4
3
2
1
0


Occupied Housing Units by Census Tract
Occupied
Housing Units
0.75
0.50
0.25
0.00

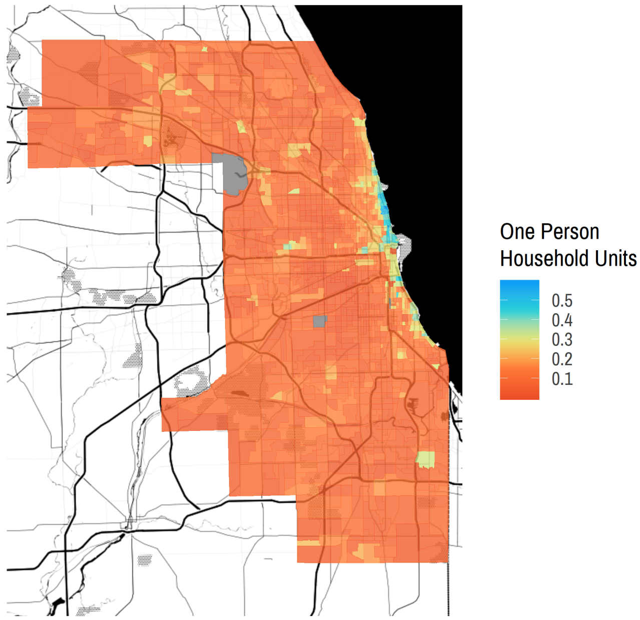

Rented Housing Units by Census Tract
Rented
Housing Units
0.75
0.50
0.25
0.00


Percent Latino by Census Tract
Percent
Latino
0.75
0.50
0.25

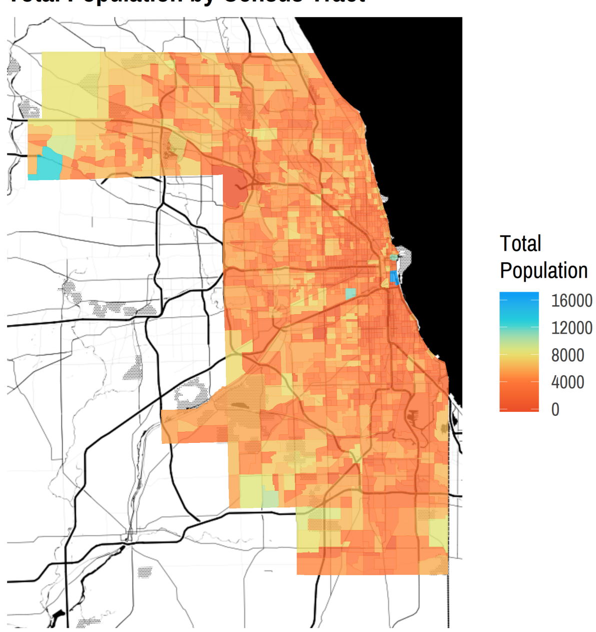

Males by Census Tract
Percent Males
0.8
0.7
0.6
0.5
0.4


Median Age by Census Tract
Median Age
60
40
20
0

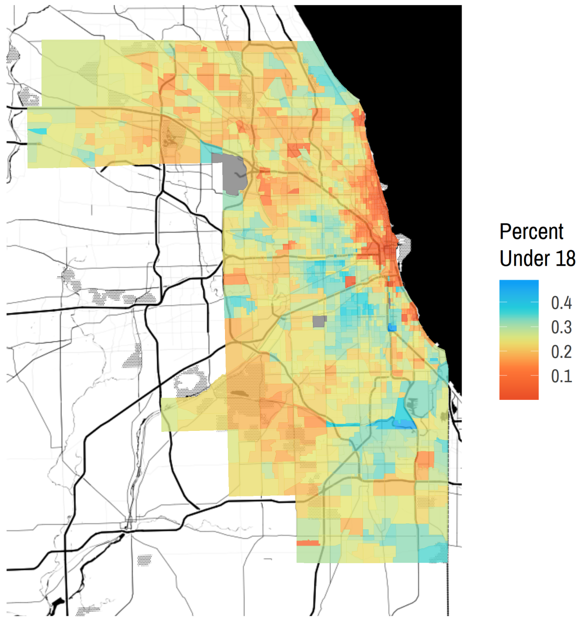
One Person Household Units by Census Tract



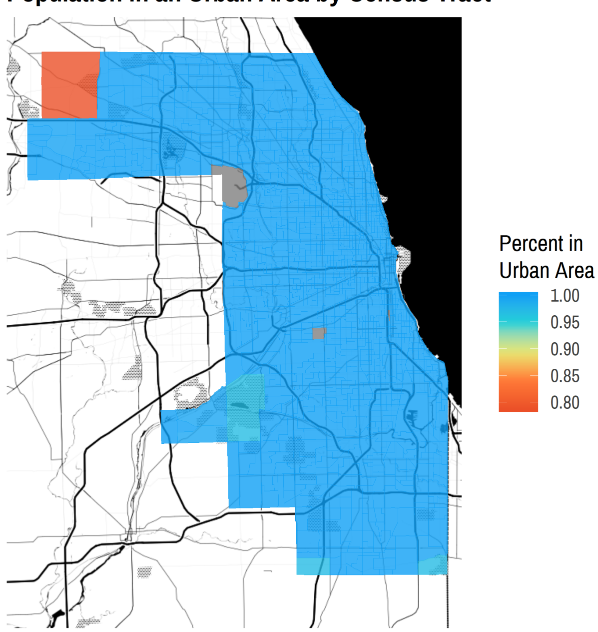
Total Population by Census Tract

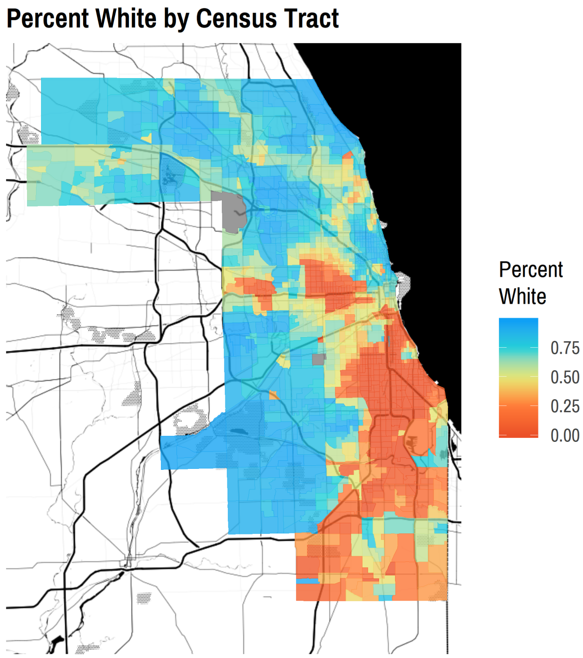


Population Under 18 by Census Tract

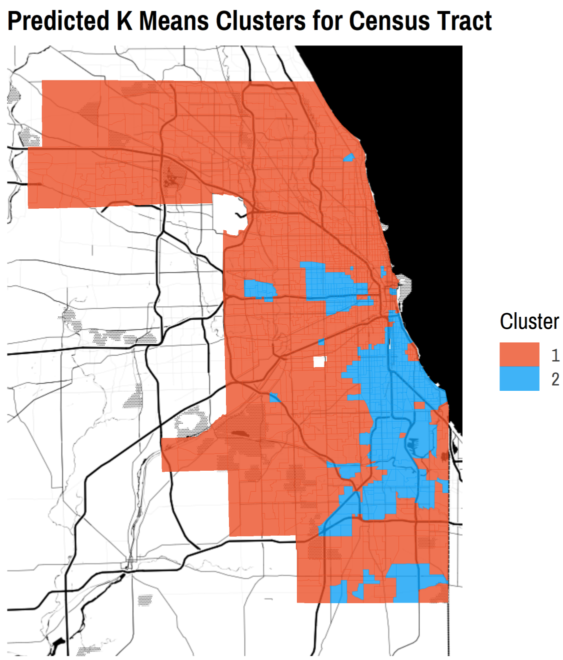


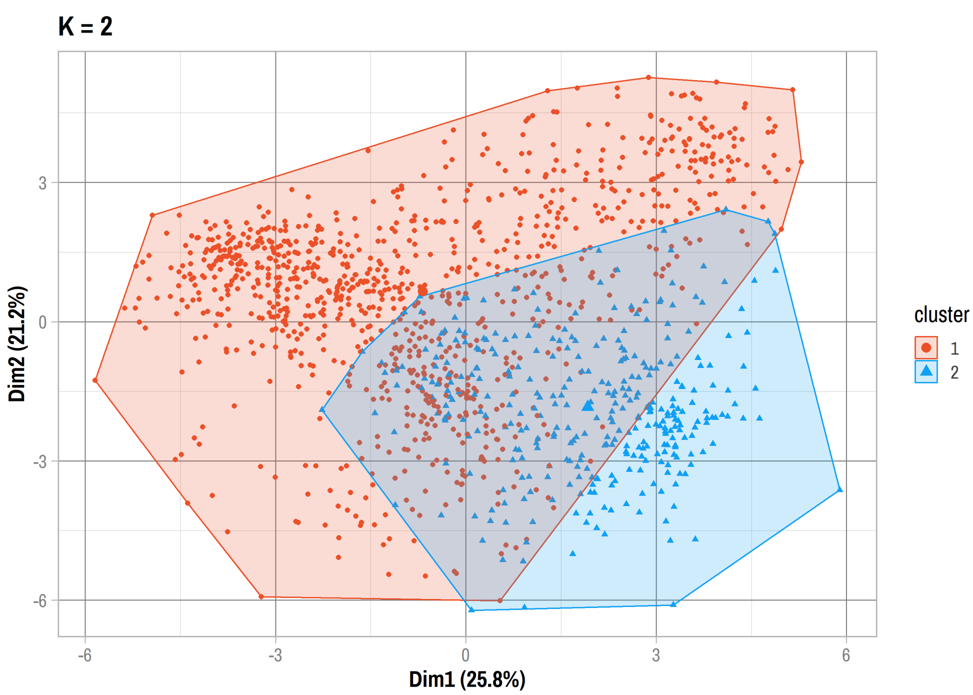
Population in an Urban Area by Census Tract

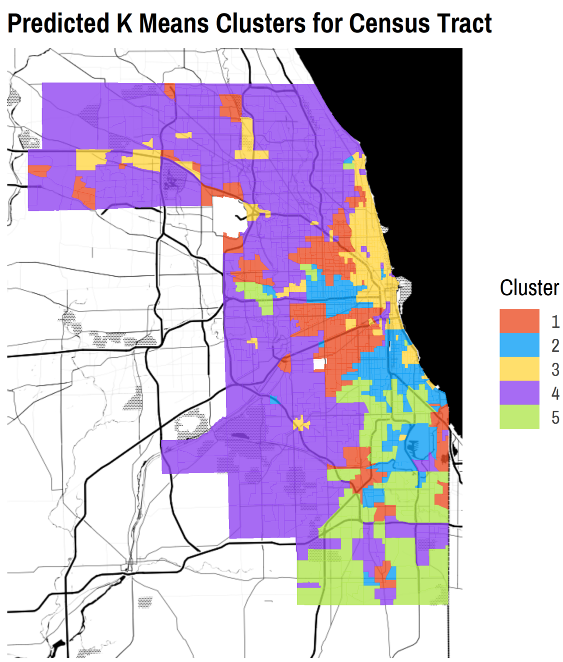


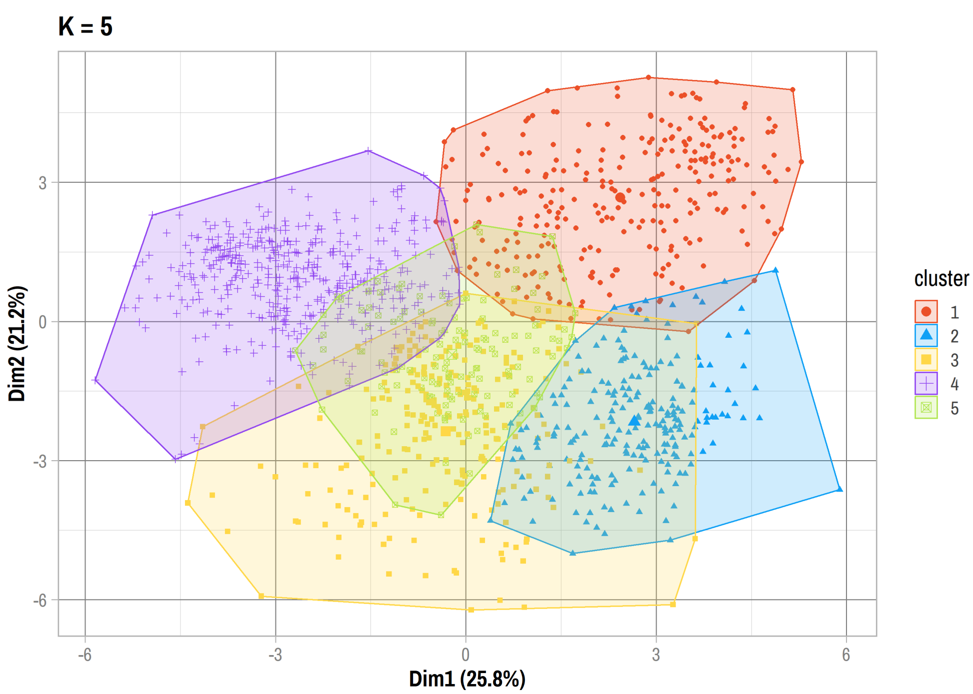

Percent White by Census Tract
Percent
White
0.75
0.50
0.25
0.00

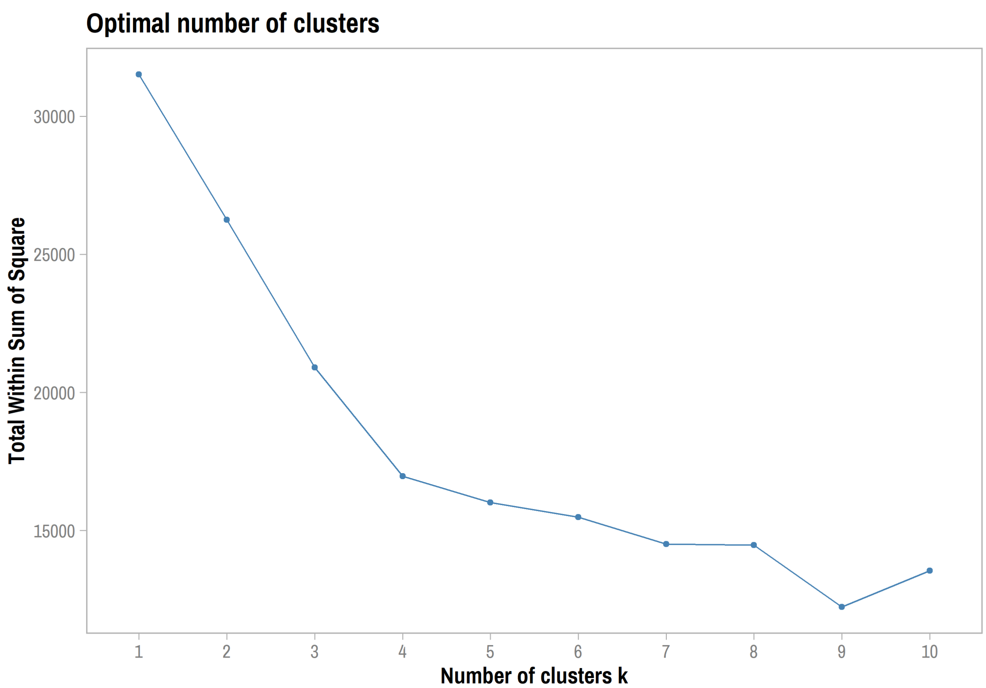
7.2 Demographic K-Means Model

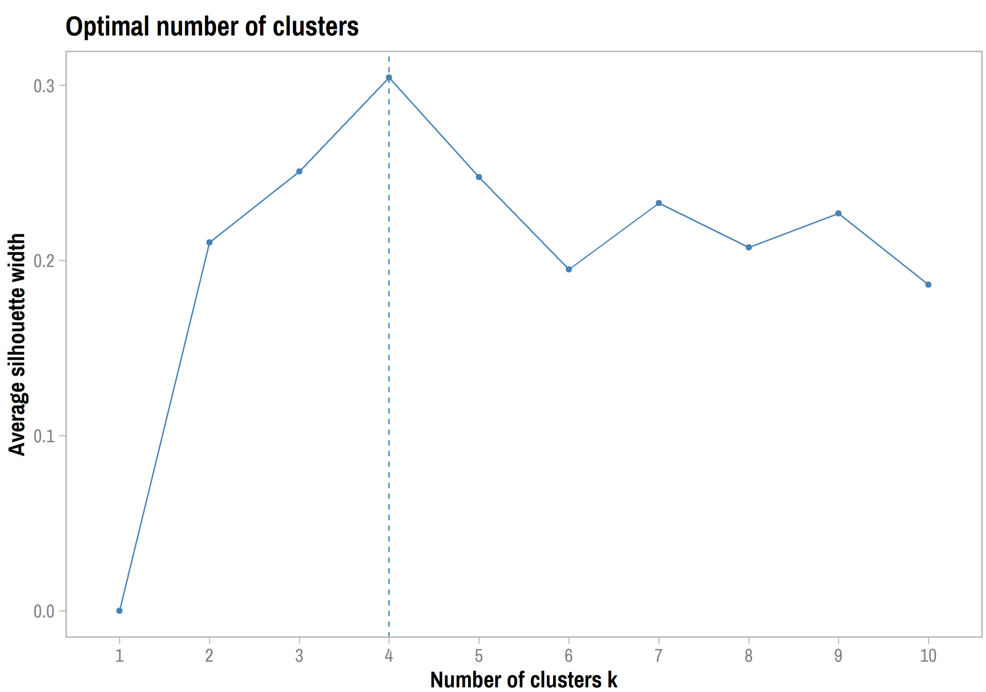


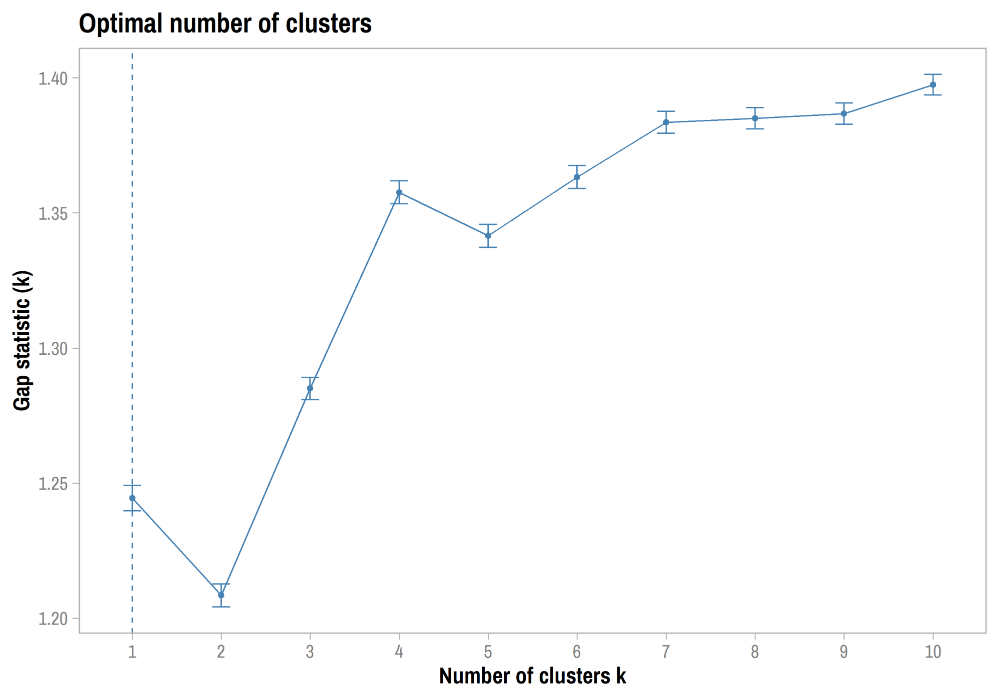


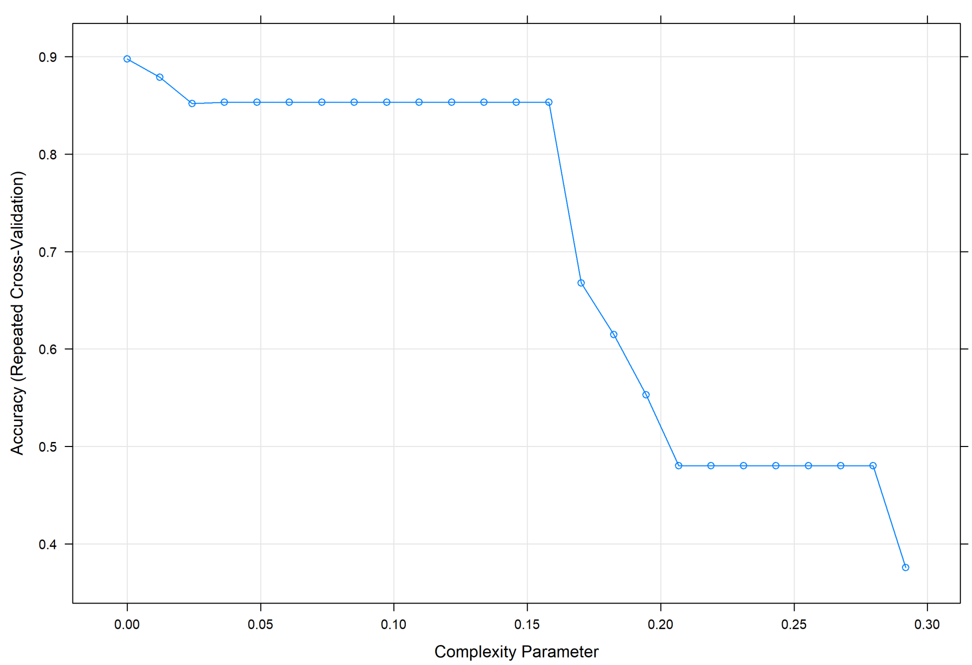

Predicted K Means Clusters for Census Tract
Cluster
1
2
3
4
5

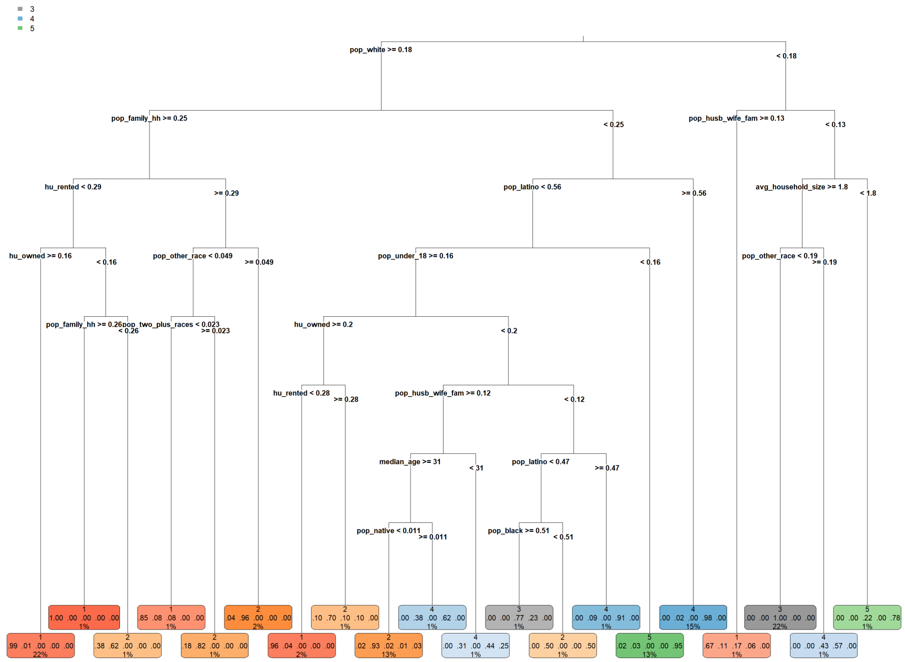

K = 5
Dim2 (21.2%)
Dim1 (25.8%)
3
0
-3
-6
-6
-3
0
3
6
cluster
1
2
3
4
5


Optimal number of clusters
Total Within Sum of Square
30000
25000
20000
15000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

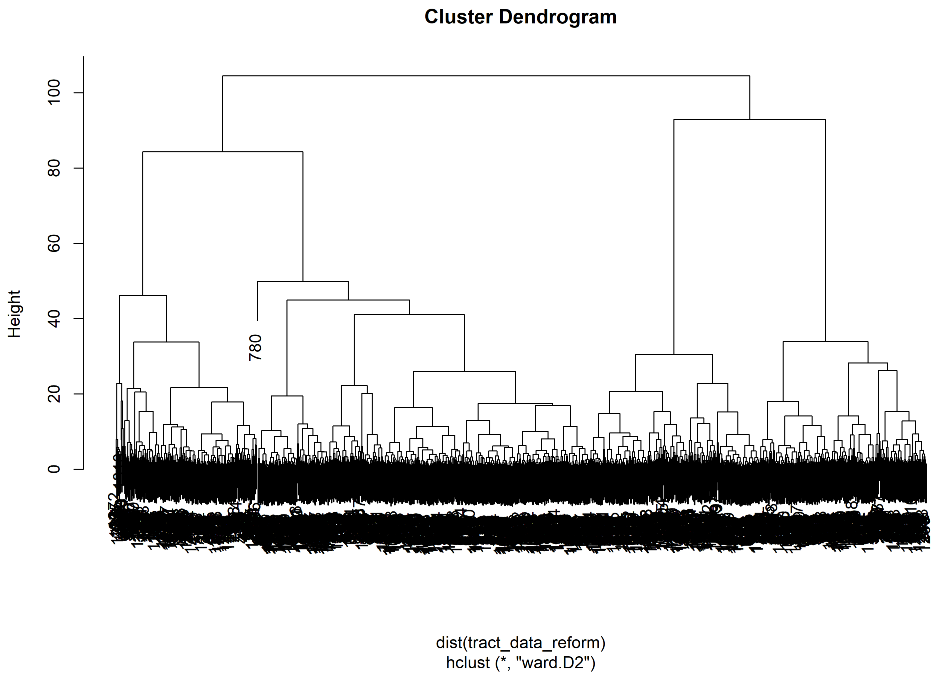

Optimal number of clusters
Average silhouette width
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
Number of clusters k

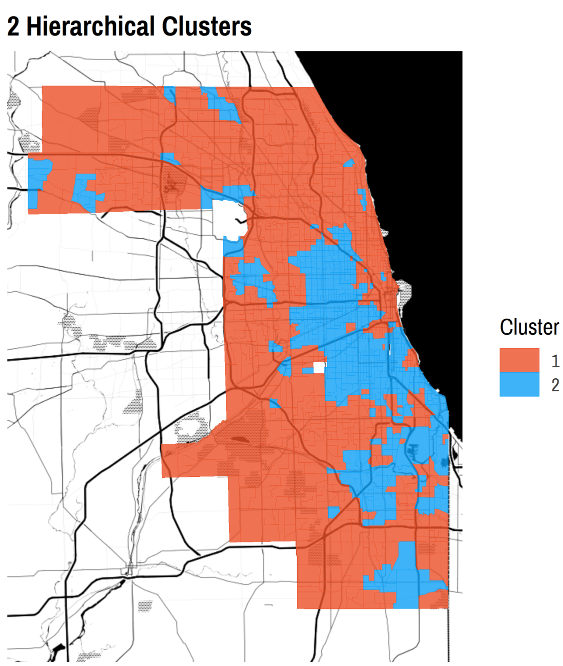

Optimal number of clusters
Gap statistic (k)
1.40
1.35
1.30
1.25
1.20
1
2
3
4
5
6
7
8
9
10
Number of clusters k

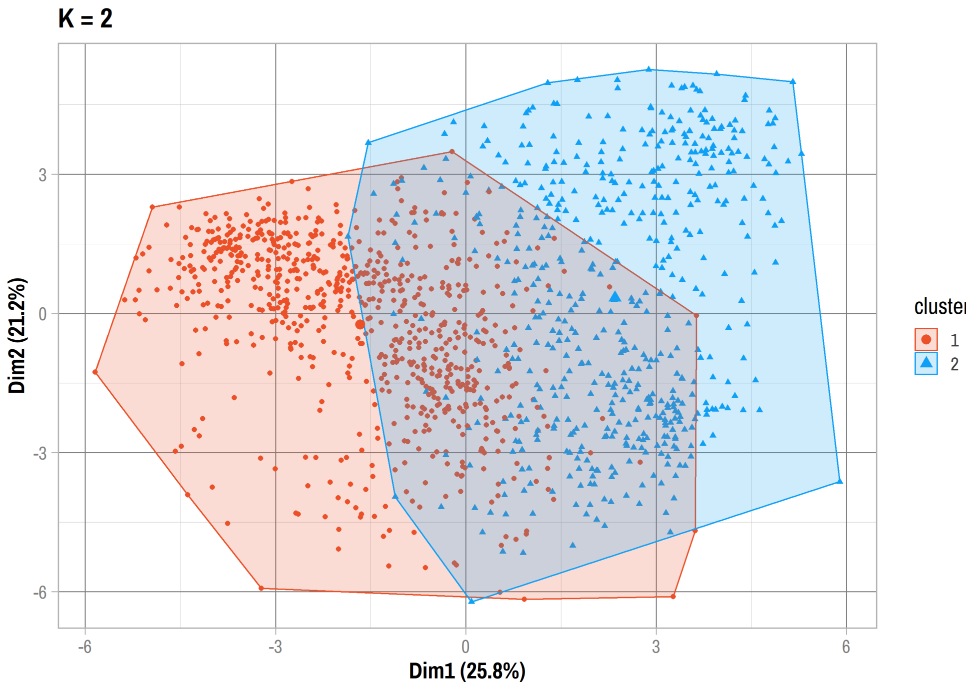

Accuracy (Repeated Cross-Validation)
0.9
0.8
0.7
0.6
0.5
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Complexity Parameter

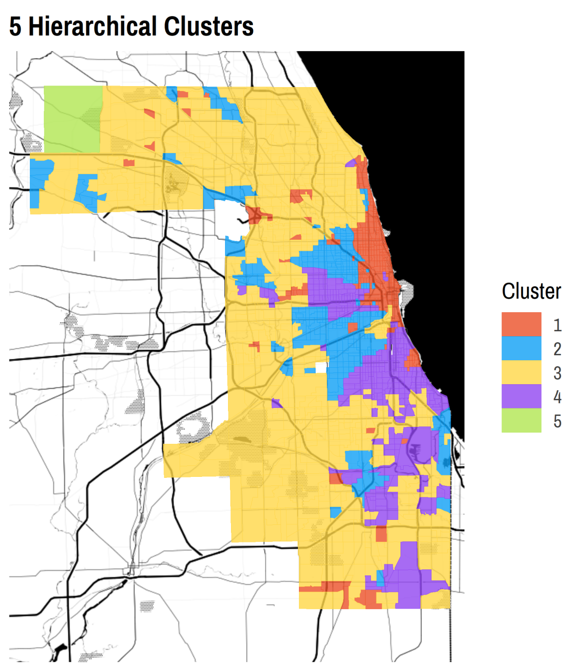

3
4
5
pop_white >= 0.18
< 0.18
pop_family_hh >= 0.25
< 0.25
pop_husb_wife_fam >= 0.13
< 0.13
hu_rented < 0.29
>= 0.29
pop_latino < 0.56
>= 0.56
avg_household_size >= 1.8
< 1.8
hu_owned >= 0.16
< 0.16
pop_other_race < 0.049
>= 0.049
pop_under_18 >= 0.16
< 0.16
pop_other_race < 0.19
>= 0.19
pop_family_hh >= 0.26
< 0.26
pop_two_plus_races < 0.023
>= 0.023
hu_owned >= 0.2
< 0.2
hu_rented < 0.28
>= 0.28
pop_husb_wife_fam >= 0.12
< 0.12
median_age >= 31
< 31
pop_latino < 0.47
>= 0.47
pop_native < 0.011
>= 0.011
pop_black >= 0.51
< 0.51
1 .99 .01 .00 .00 .00 22%
1 1.00 .00 .00 .00 .00 1%
2 .38 .62 .00 .00 .00 1%
1 .85 .08 .08 .00 .00 1%
2 .18 .82 .00 .00 .00 1%
2 .04 .96 .00 .00 .00 2%
1 .96 .04 .00 .00 .00 2%
2 .10 .70 .10 .10 .00 1%
2 .02 .93 .02 .01 .03 13%
4 .00 .38 .00 .62 .00 1%
4 .00 .31 .00 .44 .25 1%
3 .00 .00 .77 .23 .00 1%
2 .00 .50 .00 .00 .50 1%
4 .00 .09 .00 .91 .00 1%
5 .02 .03 .00 .00 .95 13%
4 .00 .02 .00 .98 .00 15%
1 .67 .11 .17 .06 .00 1%
3 .00 .00 1.00 .00 .00 22%
4 .00 .00 .43 .57 .00 1%
5 .00 .00 .22 .00 .78 1%

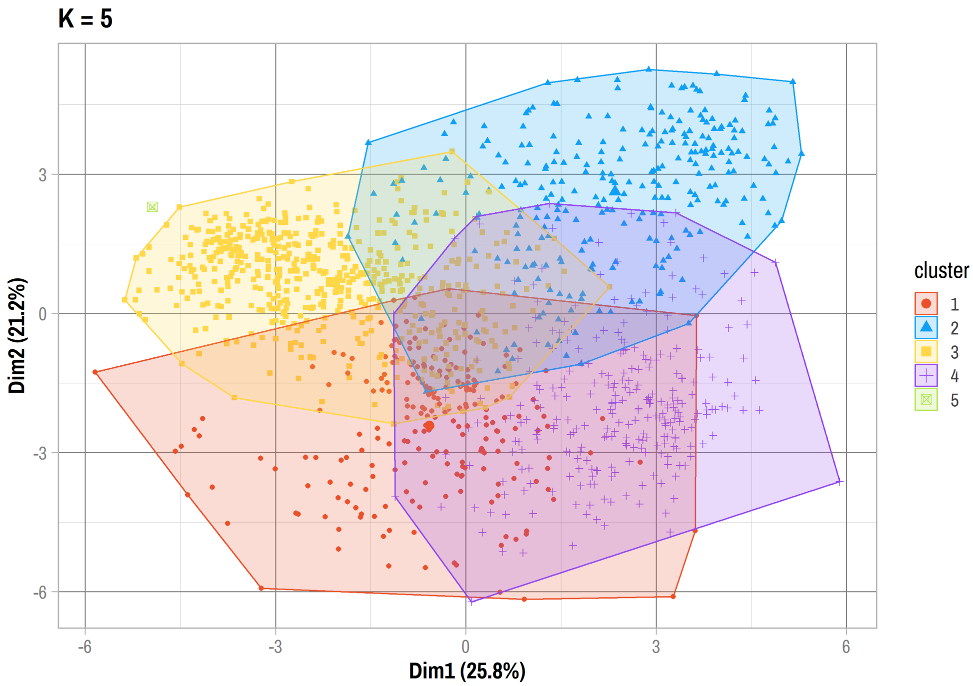
7.3 Demographic Hierarchical Model

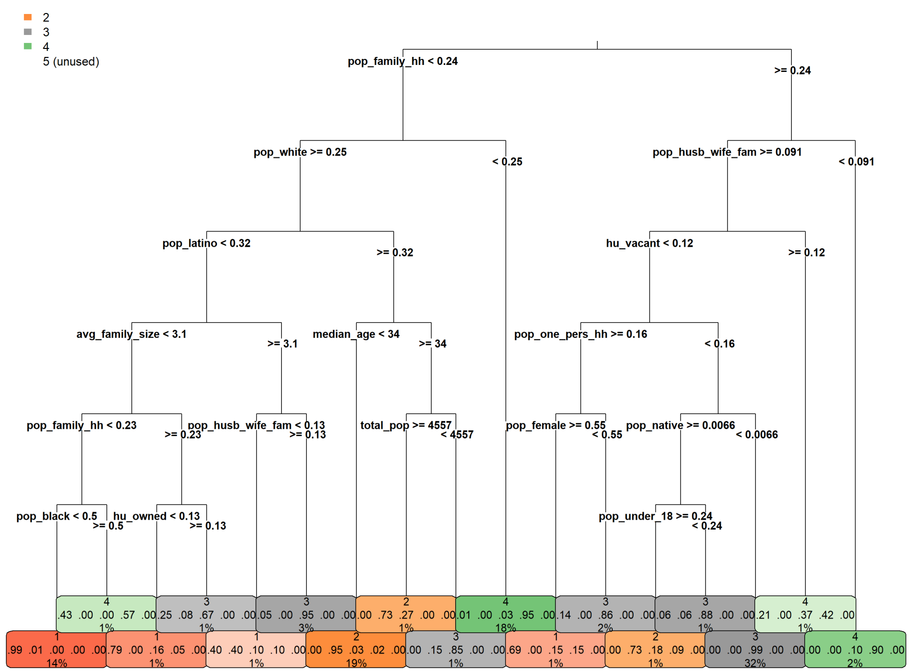




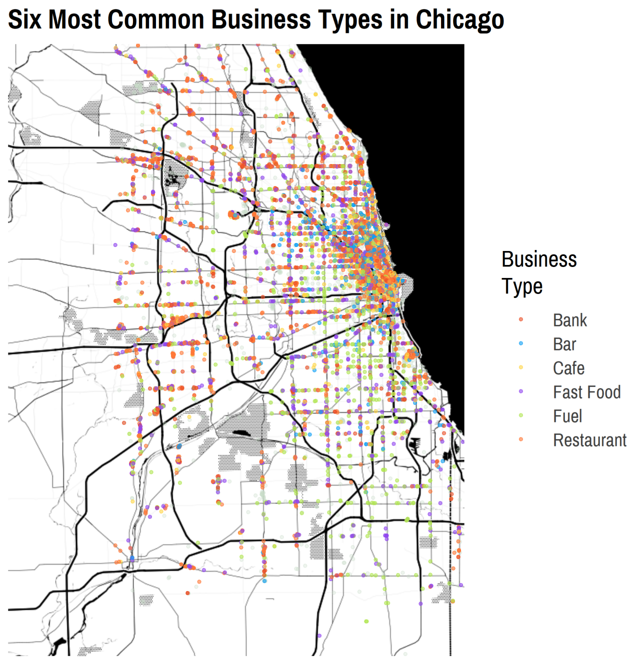

K = 2
Dim2 (21.2%)
Dim1 (25.8%)
3
0
-3
-6
-6
-3
0
3
6
cluster
1
2


5 Hierarchical Clusters
Cluster
1
2
3
4
5

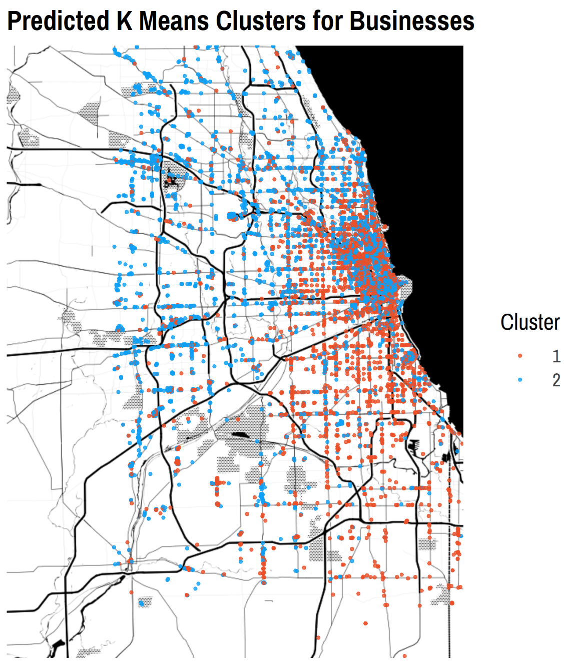

K = 5
3
0
-3
-6
Dim2 (21.2%)
-6
-3
0
3
6
Dim1 (25.8%)
cluster
1
2
3
4
5

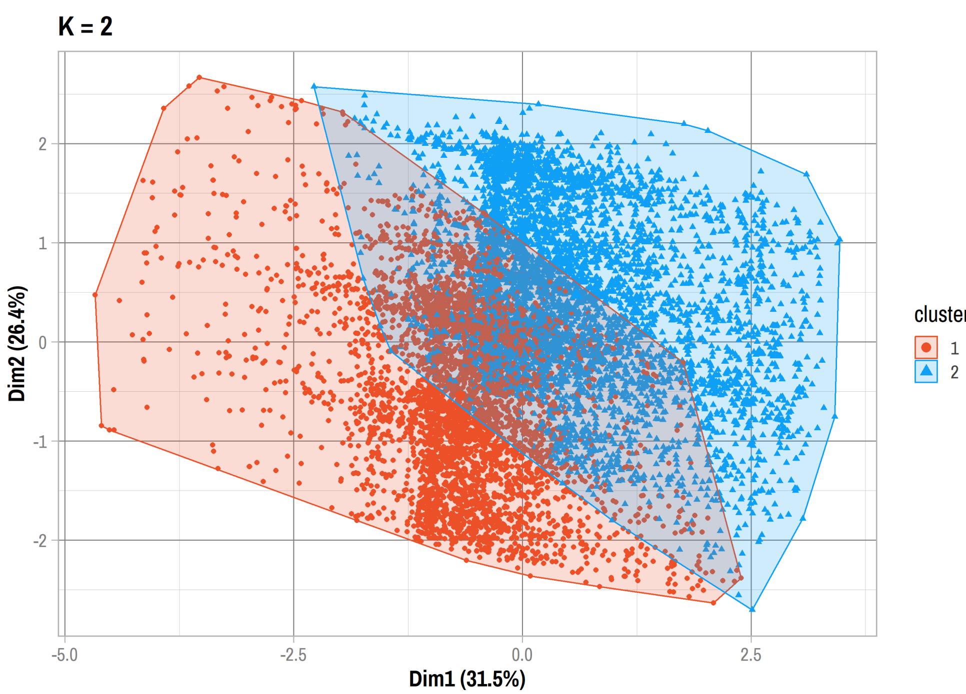

2
3
4
5 (unused)
pop_family_hh < 0.24
>= 0.24
pop_white >= 0.25
< 0.25
pop_husb_wife_fam >= 0.091
< 0.091
pop_latino < 0.32
>= 0.32
hu_vacant < 0.12
>= 0.12
avg_family_size < 3.1
>= 3.1
median_age < 34
>= 34
pop_one_pers_hh >= 0.16
< 0.16
pop_family_hh < 0.23
>= 0.23
pop_husb_wife_fam < 0.13
>= 0.13
total_pop >= 4557
< 4557
pop_female >= 0.55
< 0.55
pop_native >= 0.0066
< 0.0066
pop_black < 0.5
>= 0.5
hu_owned < 0.13
>= 0.13
pop_under_18 >= 0.24
< 0.24
4 .43 .00 .00 .57 .00 1%
3 .25 .08 .67 .00 .00 1%
3 .05 .00 .95 .00 .00 3%
2 .00 .73 .27 .00 .00 1%
4 .01 .00 .03 .95 .00 18%
3 .14 .00 .86 .00 .00 2%
3 .06 .06 .88 .00 .00 1%
4 .21 .00 .37 .42 .00 1%
1 .99 .01 .00 .00 .00 14%
1 .79 .00 .16 .05 .00 1%
1 .40 .40 .10 .10 .00 1%
2 .00 .95 .03 .02 .00 19%
3 .00 .15 .85 .00 .00 1%
1 .69 .00 .15 .15 .00 1%
2 .00 .73 .18 .09 .00 1%
3 .00 .00 .99 .00 .00 32%
4 .00 .00 .10 .90 .00 2%

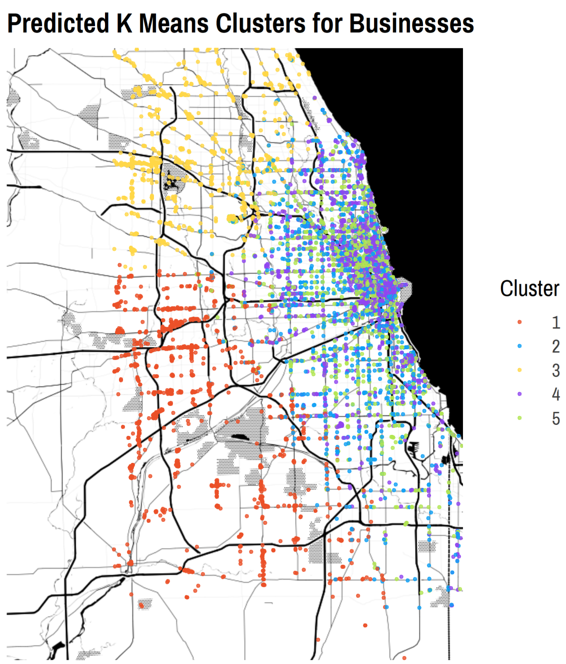
7.4 Business Data

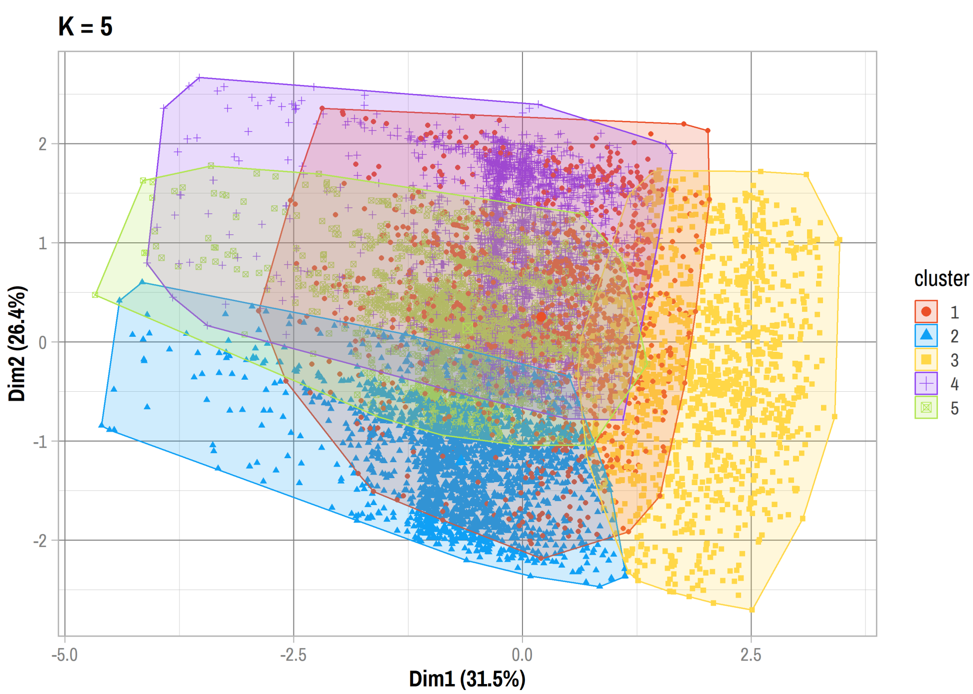


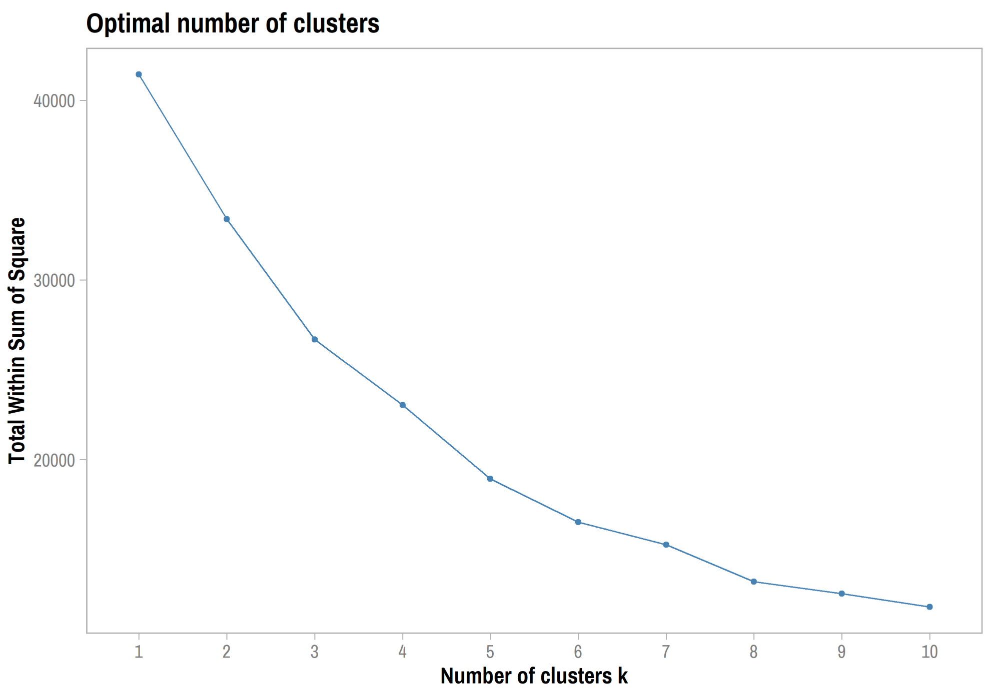
7.5 Business K-means Model

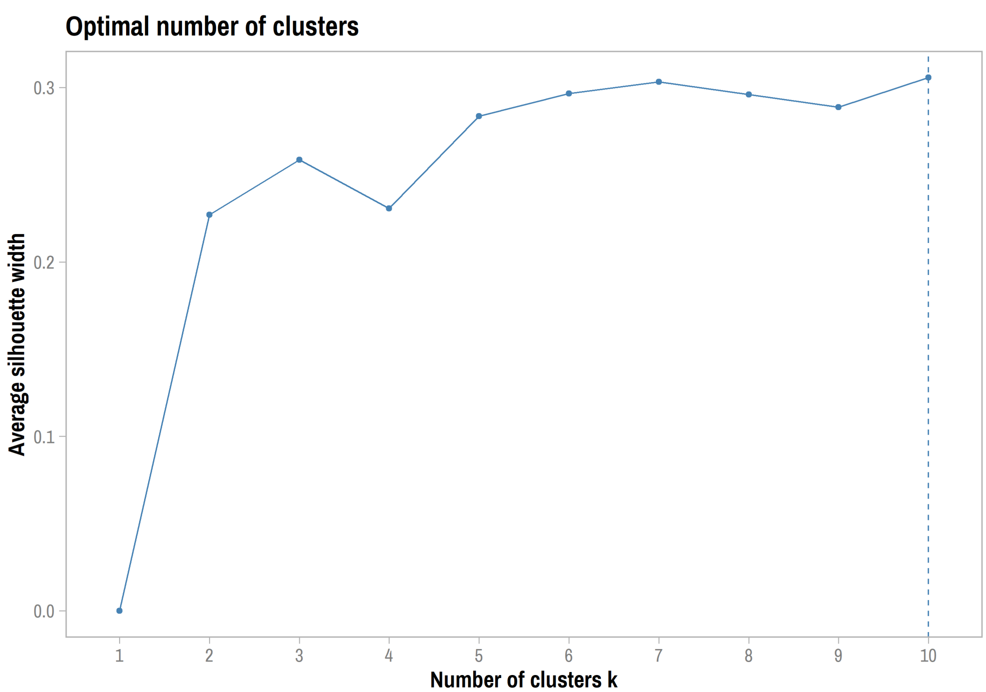




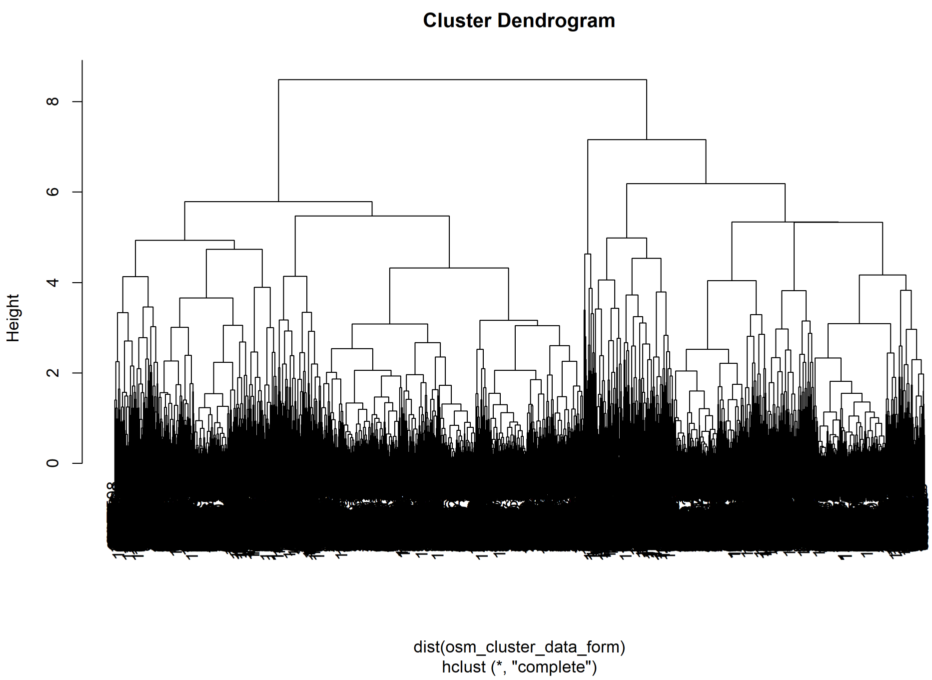

Predicted K Means Clusters for Businesses
Cluster
1
2
3
4
5

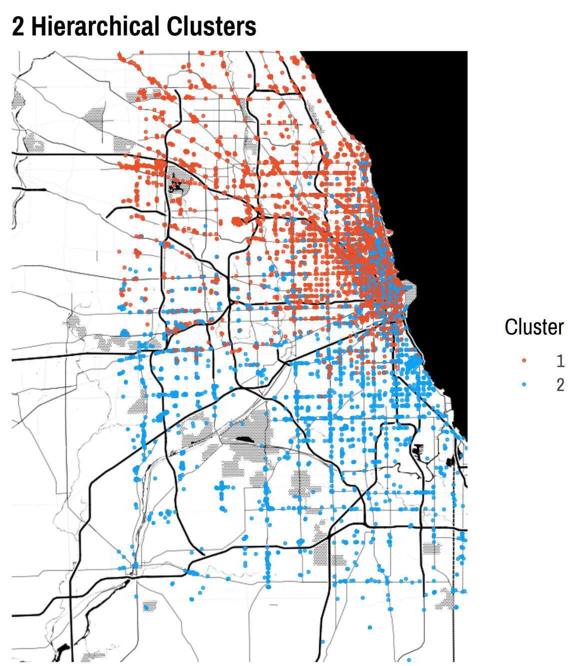

K = 5
Dim2 (26.4%)
2
1
0
-1
-2
-5.0
-2.5
0.0
2.5
Dim1 (31.5%)
cluster
1
2
3
4
5

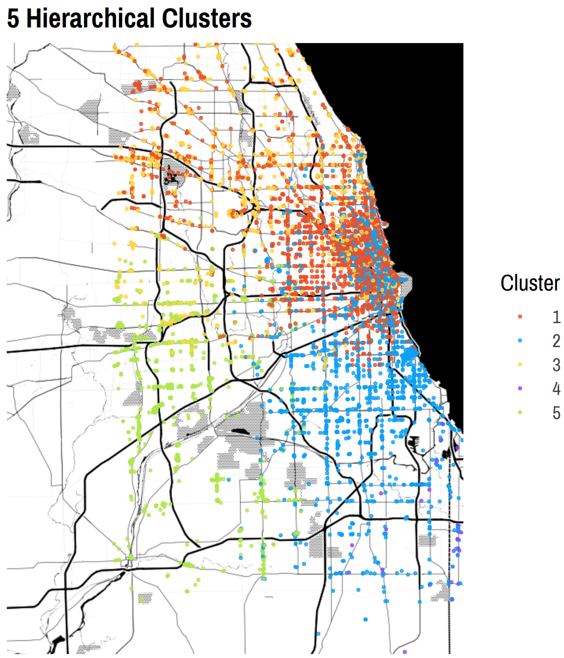

Optimal number of clusters
Total Within Sum of Square
40000
30000
20000
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
Number of clusters k

7.6 Business Hierarchical Model






5 Hierarchical Clusters
Cluster
1
2
3
4
5