Data and Methods

Adam Shelton

5/22/2019

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.

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.

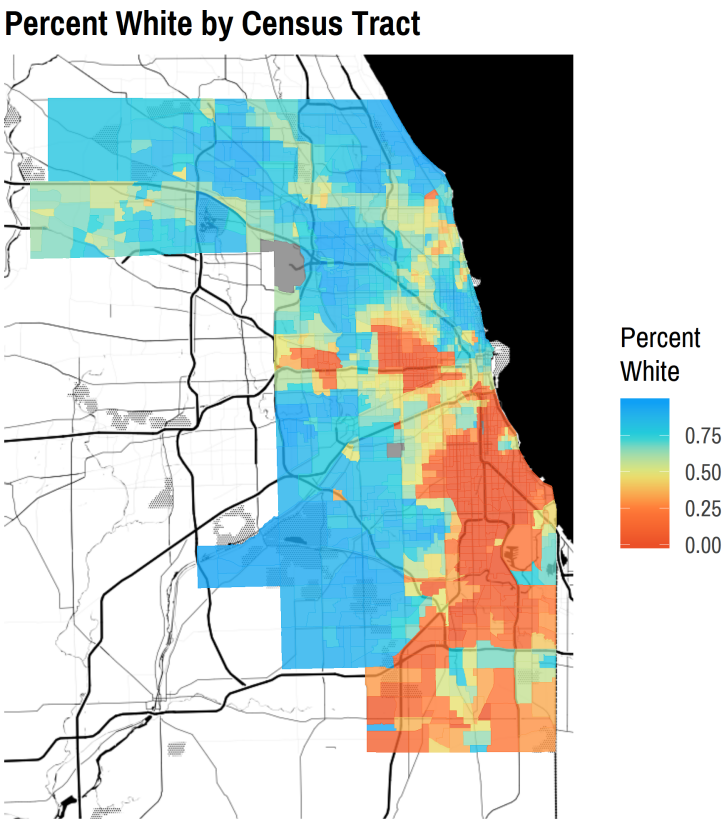
	avg_family_size	avg_household_size	median_age	total_housing_units	hu_owned	hu_owned_loan
Mean	3.314	2.681	35.663	1653.039	0.132	0.375
Stdev	0.490	0.604	6.737	842.616	0.090	0.170
Median	3.270	2.670	34.800	1559.000	0.108	0.362
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	4.730	4.390	61.300	11681.000	0.690	0.811
1. Quartile	3.040	2.310	30.800	1094.500	0.059	0.239
3. Quartile	3.600	3.040	40.300	2071.000	0.195	0.517
Variance	0.240	0.365	45.388	710001.376	0.008	0.029
Stdev.1	0.490	0.604	6.737	842.616	0.090	0.170
Skewness	-0.933	-0.158	-0.054	2.390	0.926	0.149

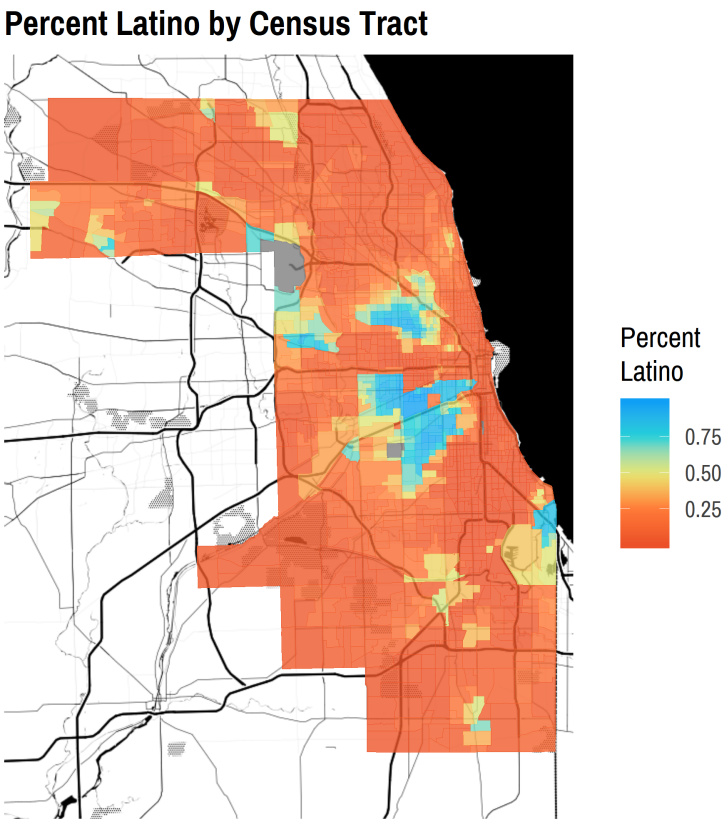
	hu_rented	hu_vacant	total_pop	pop_asian	pop_black	pop_family_hh
Mean	0.388	0.106	3938.343	0.056	0.298	0.230
Stdev	0.198	0.078	1801.869	0.084	0.376	0.038
Median	0.410	0.083	3839.000	0.026	0.061	0.233
Minimum	0.000	0.014	0.000	0.000	0.000	0.019
Maximum	0.976	1.000	16735.000	0.889	0.993	0.303
1. Quartile	0.217	0.058	2539.000	0.004	0.021	0.211
3. Quartile	0.547	0.127	5147.500	0.070	0.621	0.256
Variance	0.039	0.006	3246733.438	0.007	0.142	0.001
Stdev.1	0.198	0.078	1801.869	0.084	0.376	0.038
Skewness	-0.079	3.313	0.650	3.369	0.946	-1.220

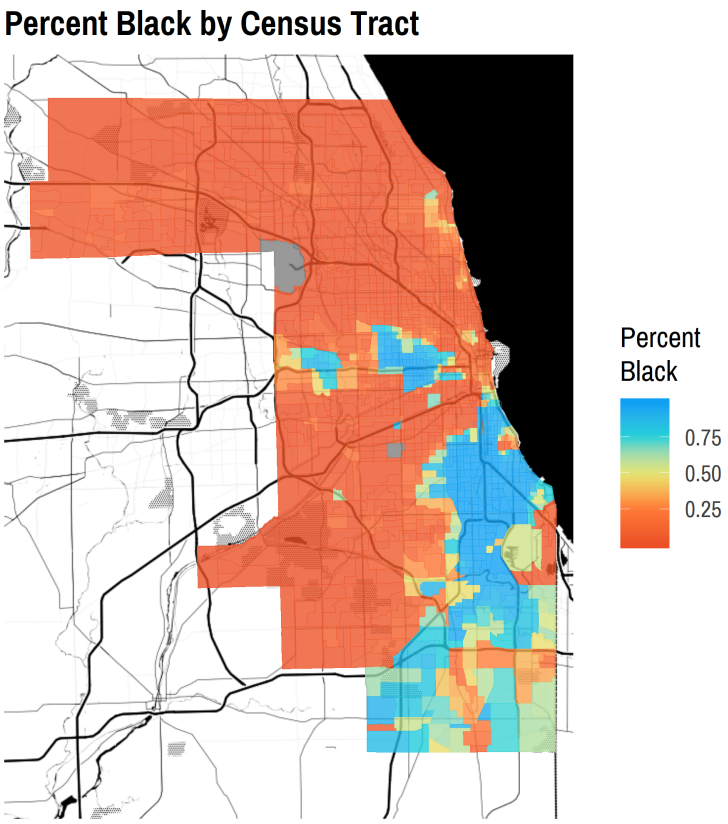
	pop_female	pop_husb_wife_fam	pop_in_hus	pop_islander	pop_latino	pop_native
Mean	0.518	0.145	0.983	0.000	0.226	0.004
Stdev	0.031	0.057	0.063	0.001	0.265	0.004
Median	0.516	0.144	0.999	0.000	0.103	0.003
Minimum	0.133	0.012	0.092	0.000	0.001	0.000
Maximum	0.654	0.274	1.000	0.008	0.987	0.026
1. Quartile	0.499	0.104	0.991	0.000	0.042	0.001
3. Quartile	0.537	0.185	1.000	0.000	0.321	0.006
Variance	0.001	0.003	0.004	0.000	0.070	0.000
Stdev.1	0.031	0.057	0.063	0.001	0.265	0.004
Skewness	-1.190	0.044	-9.130	3.962	1.398	1.886

	pop_one_pers_hh	pop_other_race	pop_rural	pop_two_plus_races	pop_under_18	pop_white
Mean	0.123	0.101	0.000	0.025	0.236	0.516
Stdev	0.087	0.132	0.006	0.012	0.074	0.321
Median	0.098	0.036	0.000	0.024	0.242	0.580
Minimum	0.003	0.000	0.000	0.001	0.013	0.000
Maximum	0.588	0.552	0.213	0.111	0.479	0.976
1. Quartile	0.067	0.011	0.000	0.015	0.198	0.206
3. Quartile	0.148	0.144	0.000	0.033	0.287	0.802
Variance	0.008	0.017	0.000	0.000	0.005	0.103
Stdev.1	0.087	0.132	0.006	0.012	0.074	0.321
Skewness	2.018	1.574	33.350	0.872	-0.581	-0.404

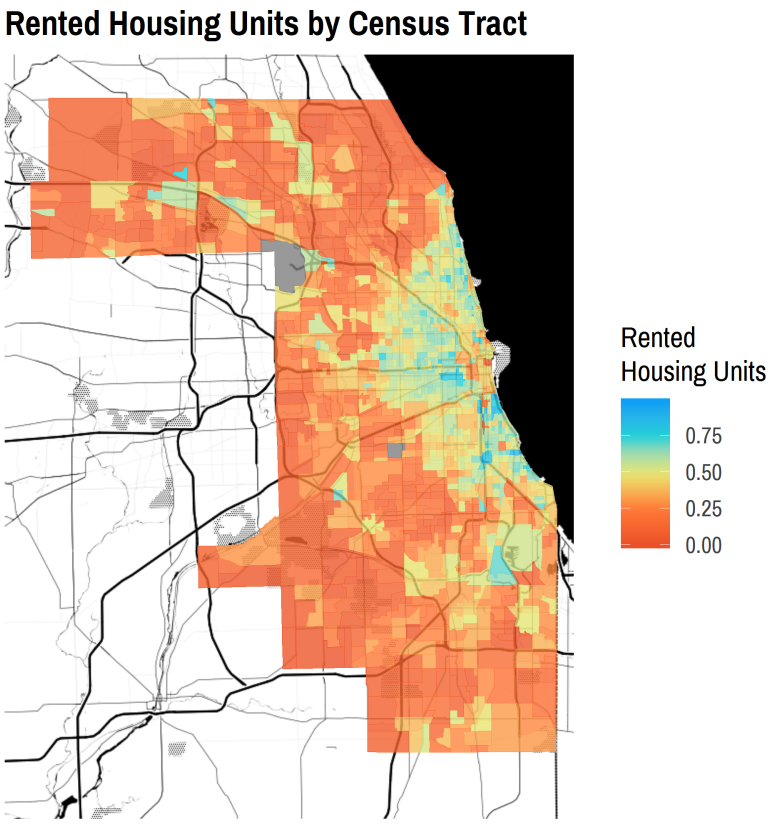
It is difficult to describe geospatial data purely by the numbers, so visualizations were generated using the `ggplot2`, `sf`, and `ggmap` packages in R. For the sake of brevity, visualizations for every variable are not included below. Chicago is notorious for its racial segregation, stemming from racialized 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.

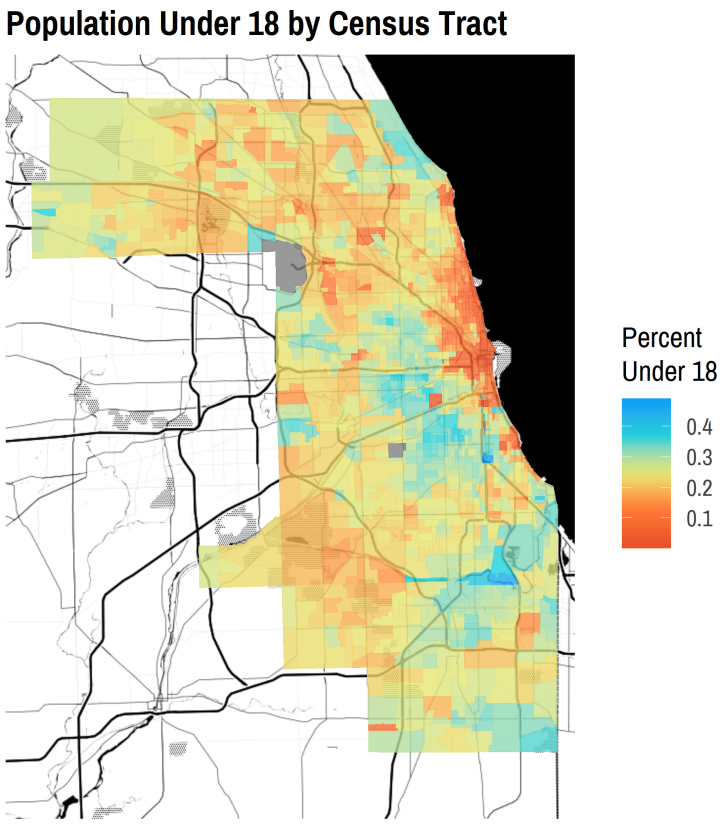

Percent White by Census Tract
Percent
White
0.75
0.50
0.25
0.00


Percent Latino by Census Tract
Percent
Latino
0.75
0.50
0.25



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.


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



We can see some of the connections between these variables with a correlation map. 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.

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

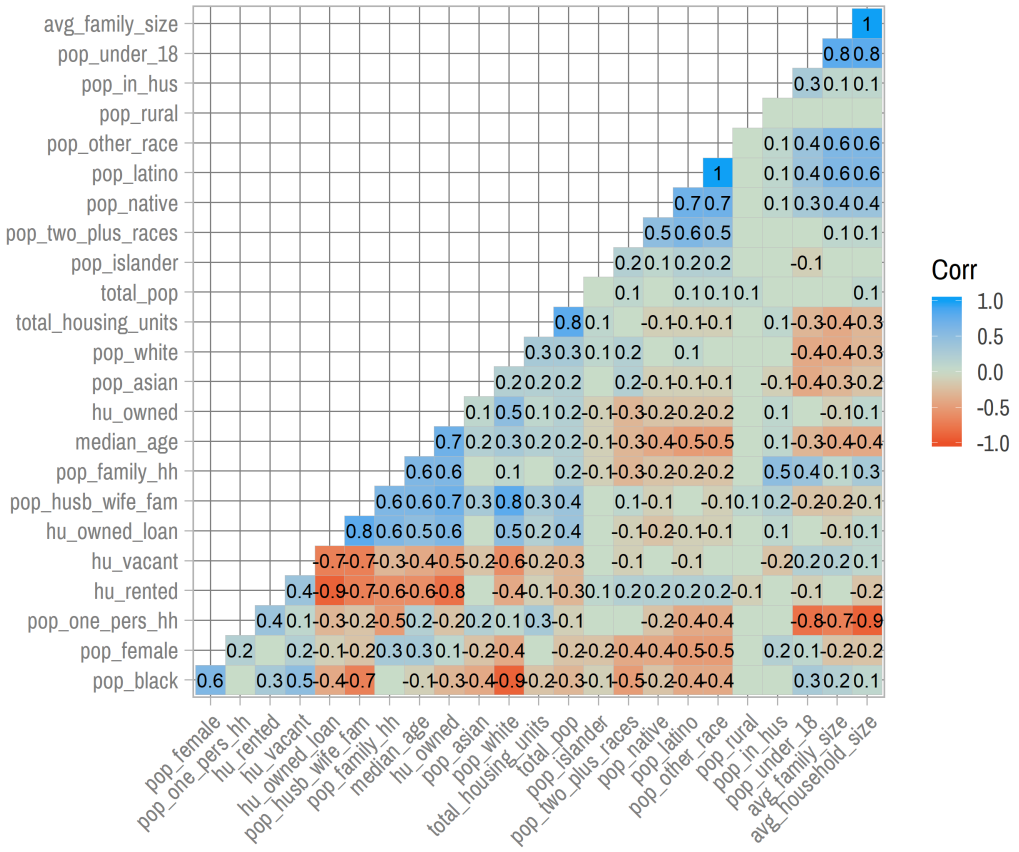


Figure 1: Demographic correlation plot

acre section of northwestern Illinois. This effectively covers all mapped 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. 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 Open Street Map's export tool.

While, there exists no pre-defined way to separate businesses from other points of interests, this can be accomplished through other means. Filtering by specific amenity categories that are wholly associated with businesses, like "banks" or "pharmacies", should provide enough means to compile a data-set of most businesses within Cook County. Other information like phone numbers or addresses could also assist with this process. While it remains unlikely that *every* business in Cook County would be in the data-set, a large enough amount of them should be captured to conduct an effective analysis.

Methods

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 of 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 Principal Component Analysis and a cross-validated decision tree surrogate model. These model analysis methods were chosen for their ease of computation and understanding.

Segregation Cluster 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 the respective functions in the R `stats` package.

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. Principal Component Analysis is also conducted on the final model, which coupled with bi-plots of the features in the model, can provide further insight into the forces driving the segregation outcome of the model.

Business Cluster 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 are the analysis methods of Principal Component Analysis and a decision tree surrogate model. However, as the Open Street Maps data consists almost exclusively of text data tied to a specific location, variables of interest that are just text must be converted to numeric factors before the data is 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.

Also, as there exists much larger amounts of missing values, especially for certain features, observations with any missing values should likely not be just dropped. More thoughtful consideration of which variables should be included, and how to accommodate missing values after that, needs to be conducted, if not perhaps even imputation of some kind.

As there is no guarantee that an unsupervised clustering model will be able to separate businesses on this information alone, further analysis may be necessary to do so. This could involve generating additional variables from text analyses conducted on the text data in the data-set, however the more productive option would likely involve a supervised clustering method, such as a Support Vector Machine. In this case, the outcome would be the predicted segregation clustering trained on the demographic data, with further models predicting specific demographic variables to gain more insight on specific types of segregation, as well as the over-all trends.

Results

Preliminary results for the demographic models are very promising, with the K-means models capturing many of the individual relationships displayed in the visual analysis of the data. These clusters clearly show the racial segregation of Chicago in addition to other factors such as age and family size/status.

This is reflected by the decision tree generated as a surrogate model. Race and ethnicity is a significant part of the model, in addition to family size/status, housing status, which reflects socioeconomic background to some degree, and median age, broken up by those younger than 31 and those under 31. Interestingly, the decision with percent white is split at 18%, with those regions being determined by family/household attributes and proportions of those who identified their race as “other”. However, contradictions to occur. For example, cluster 1 which appears to encompass tracts with a high percentage of Latinos, according the decision tree, does not even use the Latino variable in the decision process. This perhaps indicates other factors, like family size or socioeconomic class, that better represents these areas than race alone.

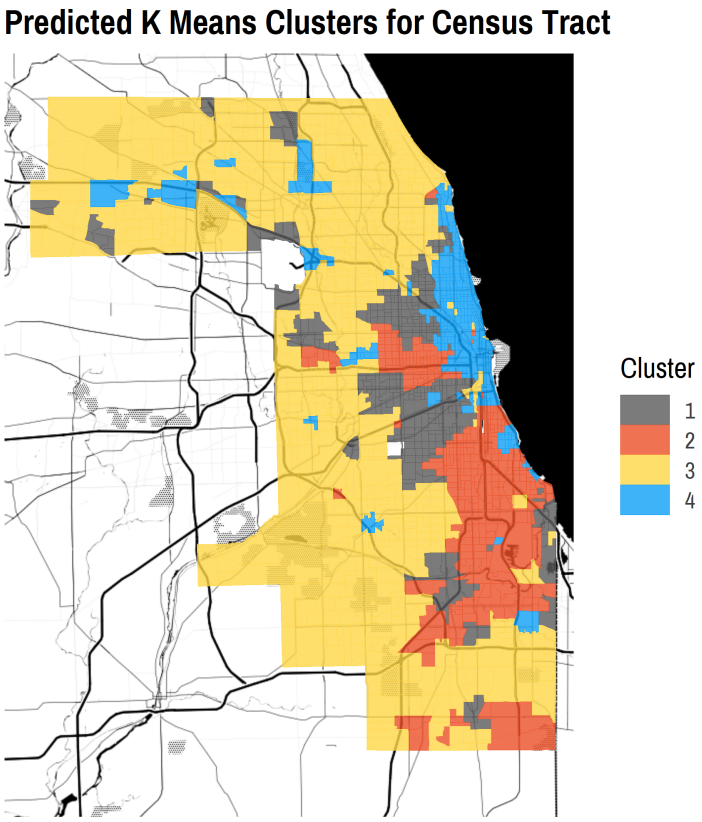


Figure 2: K-means cluster plot

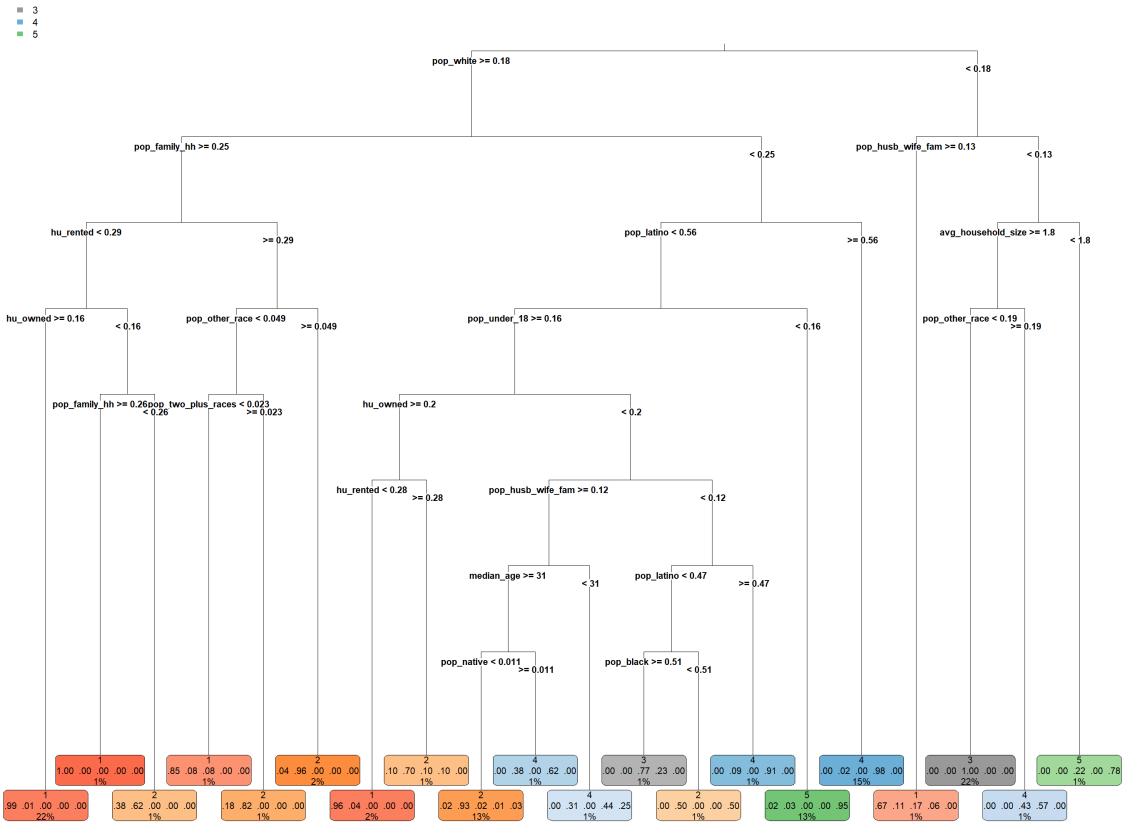


Figure 3: K-means decision tree