Algorithm development: Nicholas Marchio (Mansueto Institute for Urban Innovation)
Documentation: Nikhil Patel (Center for Spatial Data Science)
The University of Chicago
31 August 2023

How the 10 Spatial Clusters Were Created

Skew The Script (STS) is an organization that aims to provide high school teachers with math/statistics educational content that incorporate real-world data relevant to students. Lesson 4.2 on the STS website teaches students about sampling distributions. The lesson shows them a map of San Antonio (where STS is based) with 100 circles on it that represent people of different races and incomes, and students are meant to use various sampling techniques to try to determine the median income of the city. Teachers across the country have been asking STS to remake the lesson for different cities so that its data is more engaging for their students, so STS approached the UChicago Center for Spatial Data Science (CSDS) in search of a programmatic method of duplicating the lesson for other cities[1].

In developing an automated process for generating city maps for the lesson, we faced a problem: how can we split a city into exactly 10 regions such that the regions are reasonably population-balanced, roughly follow real-life demographic trends (specifically, race and income), and look compact, contiguous, and visually pleasing? All of these requirements were necessary to preserve the pedagogical value of the lesson, since two of the sampling techniques students must use depend on the 10 regions' borders.

While we could have used existing methods for regionalization (SKATER and AZP), we ended up applying a simple spatial cluster method based on the common KNN algorithm. This document outlines our implementation steps.

The Algorithm

Given tract-level income and race population data from the 2022 ACS (variable codes listed below), our method generates 10 regions that generally capture race and income gradients. Note that it is currently implemented to work with US Census tract-level data, but it would work with block-group-level and block-level data as well, although (especially income) data for smaller levels are prone to having high margins of error.

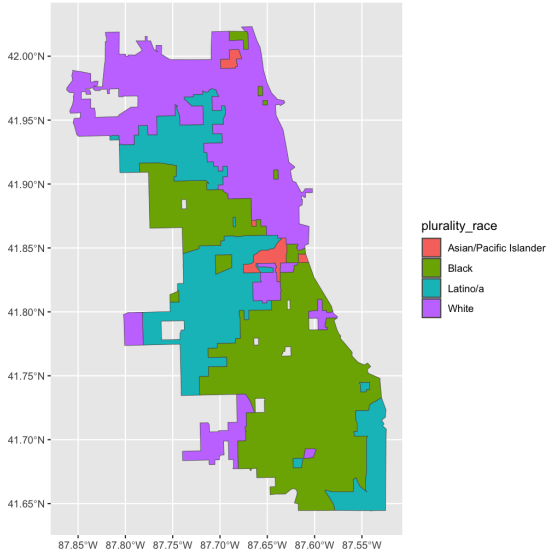
Inputs:

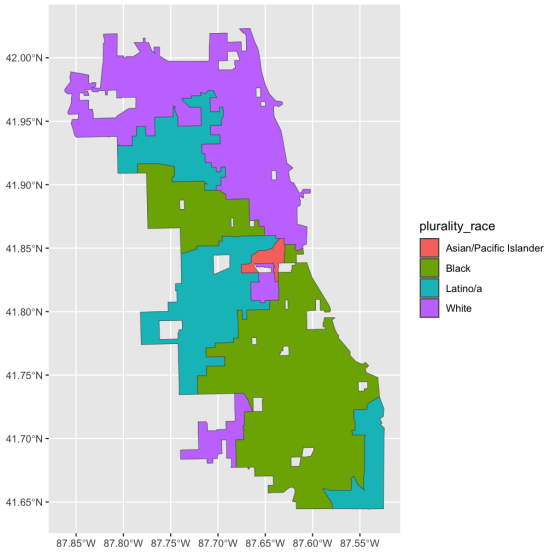
[1] The file in the GitHub repository containing the implementation of this algorithm: https://github.com/Nik4002/Geospatial-Sampling/blob/main/census-cluster-sampling-tract.R

- Geometries for the tracts that fall within the city limits, with empty tracts removed from the dataset
- Population data for each race group for each tract (the B03002 Census variable group) summarized into five groups (White, Black, Latino/a, Asian/Pacific Islander, Native American, and Multiracial/Other)
- Median income data for each tract (Census variable B19013_001)

Output: Geometries for the 10 regions

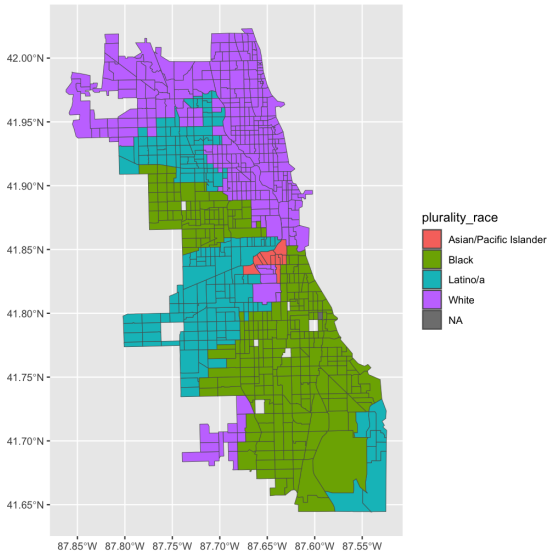
Label each tract with its plurality race and dissolve all like tracts together. In this example, this results in four non-contiguous clusters. (Lines 294-323)

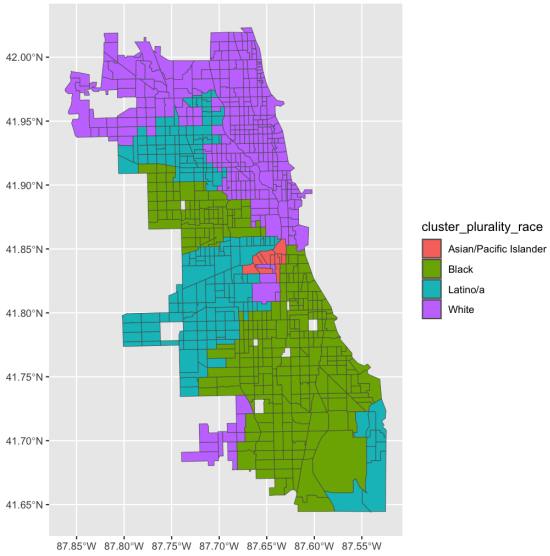




Remove any discontiguous colored regions that represent less than 1% of the city's land area. (Line 324)

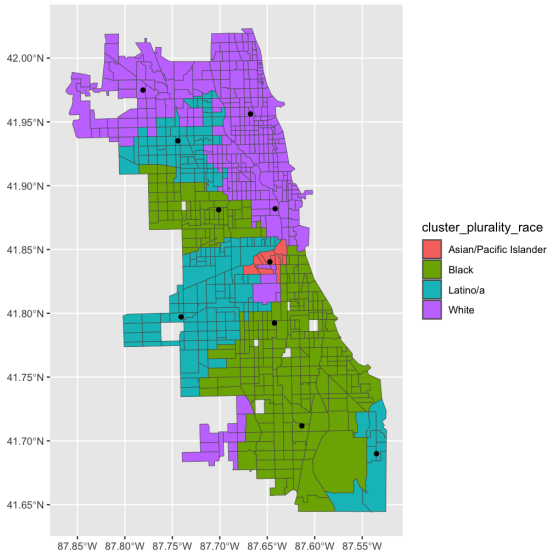
Fold removed tracts into their nearest clusters and assign each of the city's tracts a new plurality race label based on which region it falls into. At this point, we have created a smoothed-out map of the city's racial distribution. (Lines 331-338)

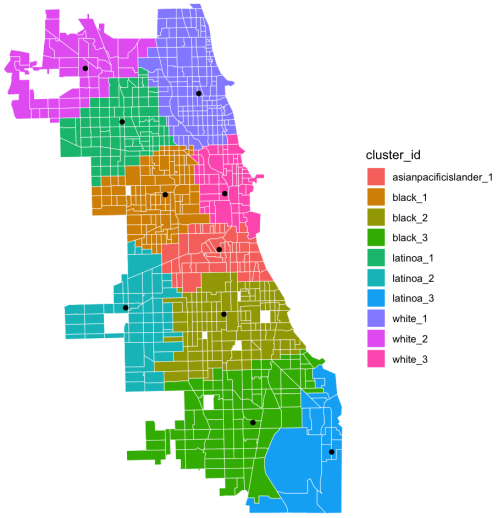




Re-dissolve and re-join tracts to fold in NA tracts (Lines 344-366)

Use the distribution of race group populations across the city to assign a total of 10 points to the different races, and use K-Means to generate points accordingly. In this example, the “white” cluster was assigned three points because Chicago’s white population is roughly 30% of its total population. (Lines 443-548)





Use KNN to reassign the city’s tracts to the ten points irrespective of each tract’s previous cluster assignment. (Lines 561-564)

Limitations and Areas for Improvement

Of the criteria we sought in a regionalization method, the one criterion we were unable to incorporate into this method yet for time reasons was strict population-balancing. To the right is a list of the 10 regions (“cluster_id”) in Chicago alongside the count of how many tracts fall within each cluster and the total population of each cluster. Optimally, Census tracts are designed to contain four

	cluster_id	count	population
	<chr>	<int>	<dbl>
1	asianpacificislander_1	54	2670608
2	black_1	107	5698819
3	black_2	123	4539819
4	black_3	91	4943107
5	latinoa_1	92	5891261
6	latinoa_2	57	3102652
7	latinoa_3	17	722243
8	white_1	147	12953831
9	white_2	41	3426984
10	white_3	55	6435873

thousand people each[2], so a method that population-balances regions on the tract level would do a reasonably good job as an overall population-balancing method. However, in practice, the population of a tract can vary from zero to over ten thousand, generally varying alongside population density (i.e. the city center often has more populous tracts than the city's outskirts). Thus, the ideal method would disregard tract counts and balance the total population of each region. Although it would be nearly impossible to guarantee a population-balanced regionalization, using tract population as a weight factor in the process would likely improve the map.

Although our method was sufficiently effective for the purpose of STS's lesson, it lacks a rigorous treatment of race and income data. It considers only each tract's plurality race, which is sufficient in many cases, especially when a city is heavily segregated. However, consider a city that has a neighborhood associated with a particular race group. If this neighborhood's tracts have a relatively high concentration of that race group but still not enough for them to qualify as the plurality group, those tracts will be misrepresented in the final result. Similarly, for diverse tracts, choosing one race group as the plurality hides the people of other races and the tracts' diversity. One possible solution to this problem is coming up with a different, more representative and comprehensive criterion for assigning race labels to tracts (instead of simply using the plurality). This method could be integrated seamlessly into the rest of the process and would thus require less effort to implement. A more rigorous solution would be to consider each tract's race distribution in creating the clusters shown early in the regionalization algorithm. For example, one could take all pairs of adjacent tracts, compare their race population distributions using some measure of distribution divergence, and have some threshold for divergence below which adjacent tracts are dissolved together into clusters.

[2] The US Census' Bureau's glossary defines census tracts: https://www.census.gov/programs-surveys/geography/about/glossary.html