Chicago Transit Equity Analysis
Using Geospatial Clustering Methods
Final Project for GIS-30519

Megan Moore

1 Project Overview

1.1 Background

Chicago has one of the busiest and most utilized public transit systems in the United States, providing 243.5 million rides in 2022 to a service community of 3.2 million people. And with an operating budget of nearly $2 billion, it has a responsibility to serve Chicagoans effectively and equitably. Transit access has a huge impact on social mobility and access to opportunity (employment, education, social connection, etc) within a city and a lack of transit can have lasting impact on people's access to resources and opportunity. Additionally, there is strong evidence that the impact of redlining has lasting impacts today, and it is important to understand if transit is a victim of this perpetuation of inequities in Chicago.

1.2 Research Question

How does transit access vary across census tracts throughout Chicago and do patterns of poor transit echo patterns of other inequities? This is broken down into three components:

1. How does bus stop location density differ across Chicago census tracts?
   (a) Since census tracts are drawn with relatively consistent population counts, we will assume that bus stop density per census tract is a comparable measure.
2. For each bus stop, how much service is scheduled for that point location in a given week? What kind of disparities do we see between census tract clusters in terms of aggregate scheduled service per tract per week?
   (a) Disparities will be considered in terms of scheduled service within a given census tract.
3. Incorporating census tract demographic information, how do transit disparities map to indicators of economic and resource disparities.
   (a) These demographic indicators will include education, employment and income data to signal areas in which poor transit service may exacerbate existing inequities.

1.3 Hypotheses

1. **Null hypothesis**: intensity of bus stop points is the same across all census tracts. → **Hypothesis rejection**: census tracts have significantly different intensities of bus stops
2. **Null hypothesis**: Scheduled bus service per census tract is the same across all tracts → **Hypothesis rejection**: there is a statistically significant difference in bus service between census tracts (by count of busses servicing census tract stops)
3. **Null hypothesis**: There is no statistically significant correlation between opportunity indicators (highest level of education, unemployment, income) and transit service (if we see a variation in service from hypotheses 1 and 2) → **Hypothesis rejection**: there is a significant relationship between opportunity indicators and poor transit service.

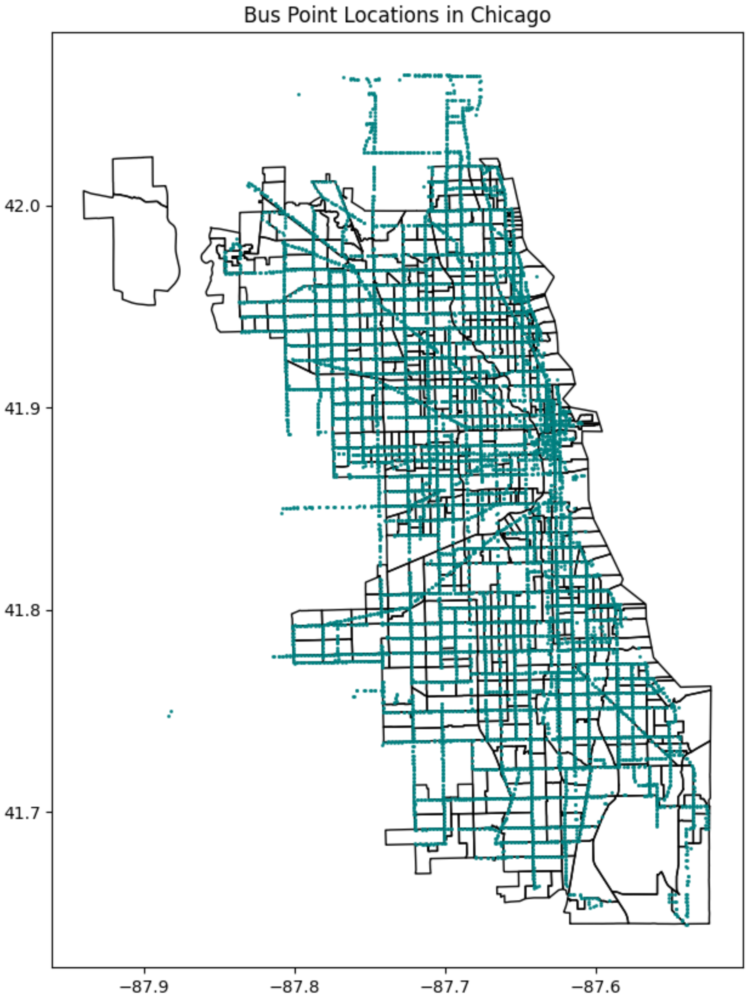
2 Data

- The main source of data is the CTA Bus Stop point data that contains all bus stops in Chicago. There are **10,760** bus stops recorded by the CTA.
- The second important source of data is the Chicago census tract data that contains census tract boundaries and other levels of data such as community area. There are **801** census tracts in the Chicago area.
- To assess service, the CTA GTFS feed is used to gather scheduled bus service to each bus stop location each week. There are **2,663,183** schedule bus service stops per week.
- To compare to sociodemographic factors, we use the Census and Opportunity atlas datasets that compile factors such as household income, and education level across different geographic granularities (counties, tracts, etc.) and for different demographic subgroups (black, white, female, male, etc.).
  1. Median household income in Chicago is **$44k**
  2. Median high school graduation rate is **86%** (this is aggregated for the county so cannot be compared on a census tract basis)
  3. Median job growth rate from 2004-2013 in Chicago is **-0.3%**
  4. Median incarceration rate is **1.2%**

3 Project Implementation

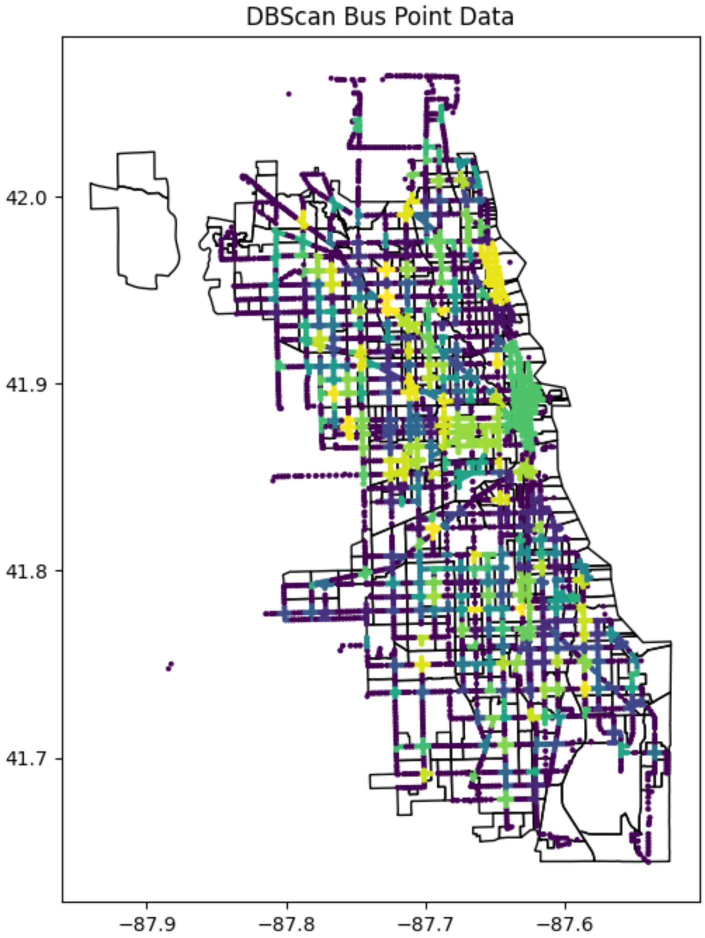
3.1 Point Data Analysis

To get an initial sense of bus stop locations throughout Chicago, I loaded and plotted the individual bus stops in point form:



3.1.1 DBScan

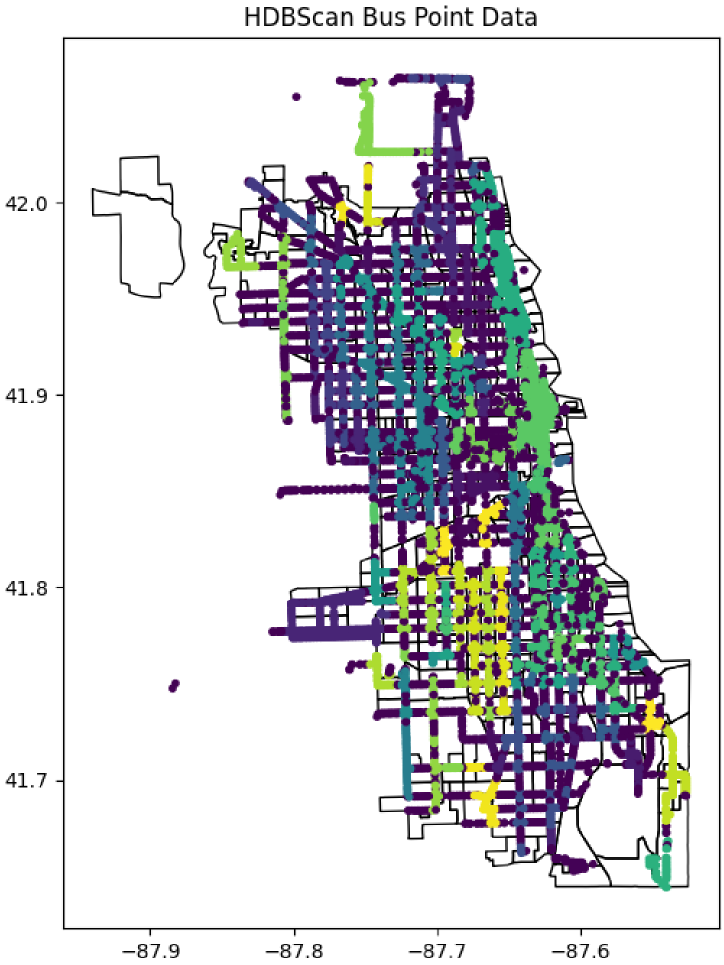
To determine how far apart bus stops are, I used HDBScan and DBScan in the hopes of being able to parse apart the areas of Chicago where there were the most nearby stops. For Density based clustering (DBScan), I found that an epsilon value of 0.003 and a minimum sample size of 10 formed the most interpretable clusters.



We see that the biggest cluster (with the most density) from this method is in the loop area, which is to be expected. There are also quite a few clusters in the North and a few smaller clusters in the South.

3.1.2 HDBScan

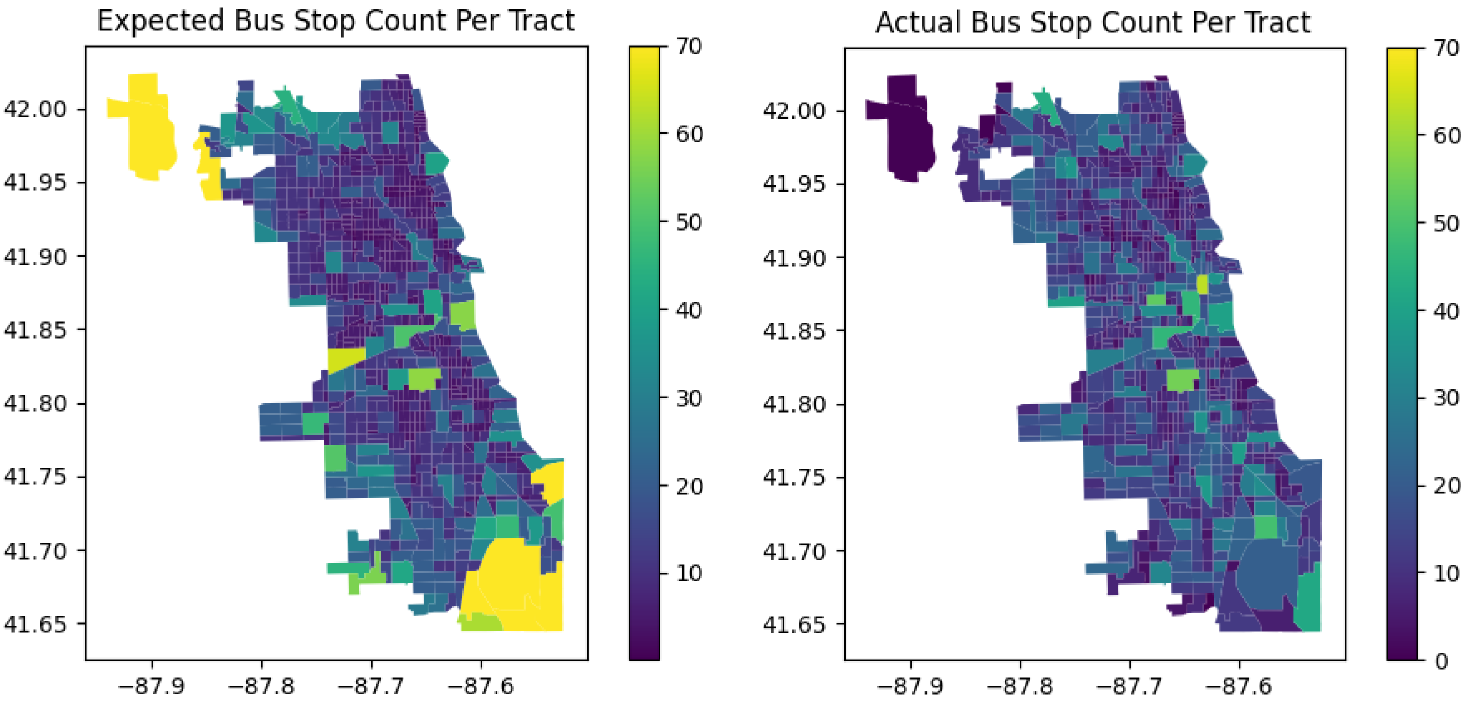
For HDBScan the minimum cluster size that worked best was 8, and the model was rather finicky making very few clusters if using many more points than that (only 2 cluster if min size was 9), and producing 351 clusters with 8. From this we got the following clusters:



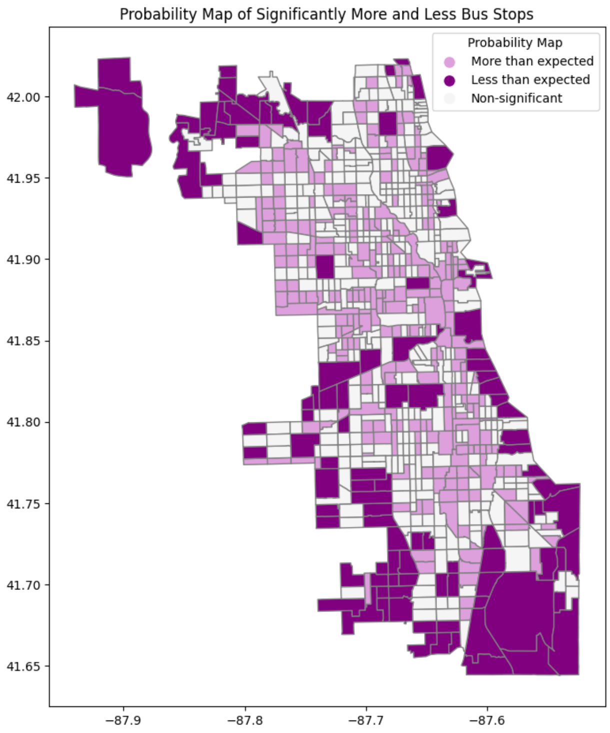
This seems to do a good job of recognizing the density around main arteries of Chicago (Milwaukee, Lakeshore drive) and reflecting on the HDBScan method, it seems to cut through the noise that may incorrectly cluster some observations in the DBScan results. Therefore this seems to be a clearer image of overarching transit density patterns throughout the city. That being said, the "noise" in DBScan may actually be useful when looking at a very granular level to understand block-by-block transit gaps.

3.2 Clustering Rates

To begin understanding the density of bus stops throughout Chicago, I aggregated the number of bus stops per census tract. I then computed the intensity of bus stops and compared the expected number of bus stops against the actual bus stop count per tract.



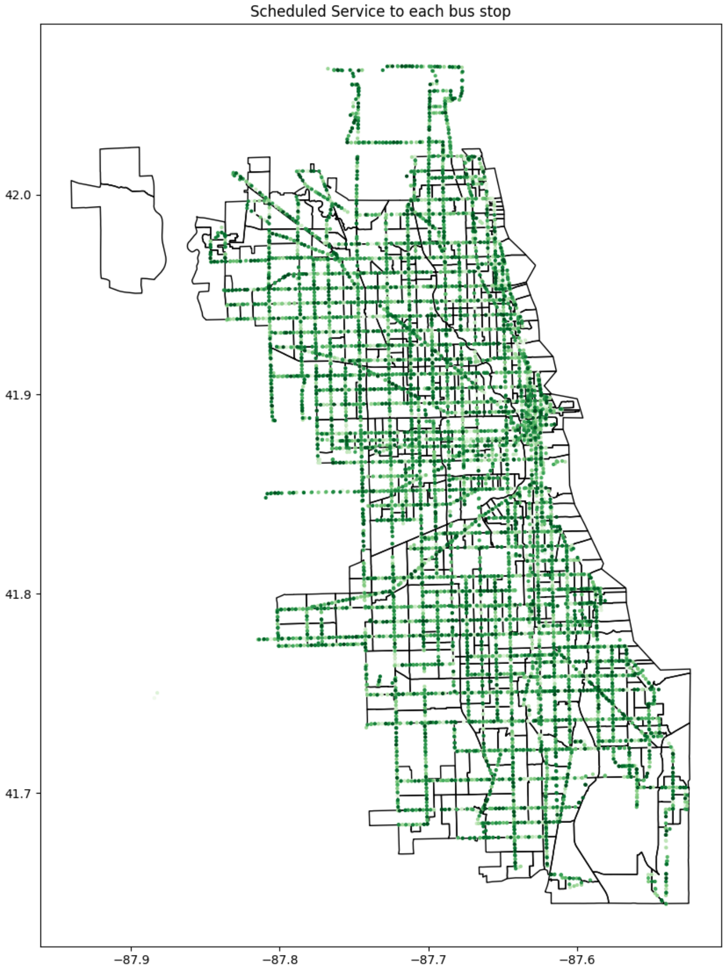
From this we can see that the large areas that may lack population density (like O'hare) are expected to have many more busses in their area than other parts of the city. Aside from that, it seems the actual counts show slightly more than expected (likely the offsetting of the initial over-expectation for the large land areas with low population density). To assess the significance of the variation of actual bus stops compared to the expected value we plot a probability map with an alpha value cutoff of 0.05.



From this we find that there are many census tracts that have significantly different bus stop counts with more bus stops than expected in the main arteries of Chicago and the further out, West and South tracts containing fewer than expected, as well as the lakefront.

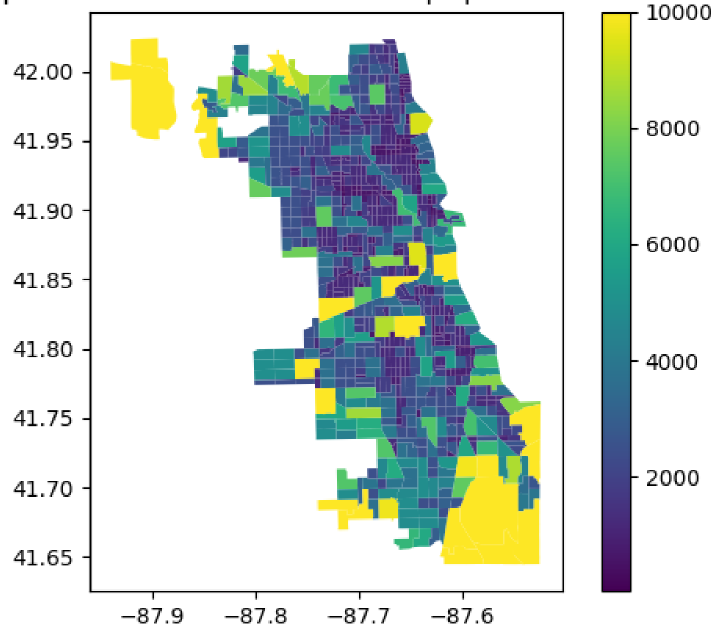
3.3 Service Density Analysis

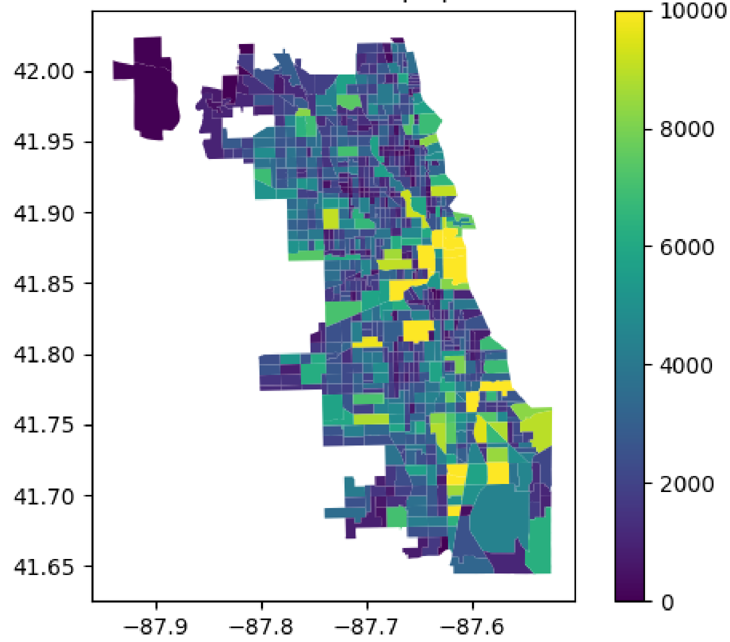
The next metric to explore on a census tract level was the scheduled service per bus stop, and subsequently the scheduled service to the bus stops in a given census tract. The hope for this analysis was to get a more nuanced perspective of what the actual access to transit in a given area may look like. After incorporating the weekly bus route schedule for each bus stop, I first plotted the bus stop point data this time using a color gradient to determine if there were specific areas that saw a higher concentration of service:



From this, there did not seem to be any specific area that had more service per stop than other areas, however this did not account for how many stops were in a given tract.

To do this I computed the service rates per census tract. First I computed the intensity of service and graphed the expected service per tract and compared it against the actual service in each census tract:

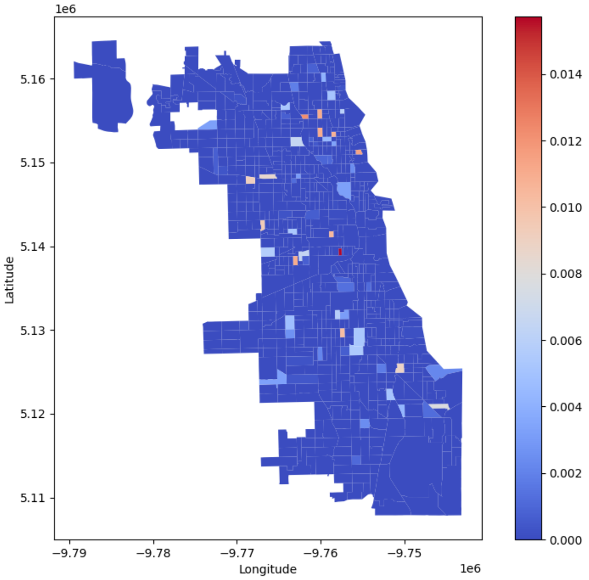

Expected Scheduled Service to Stops per Census Tract


Real Scheduled Service to Stops per Census Tract

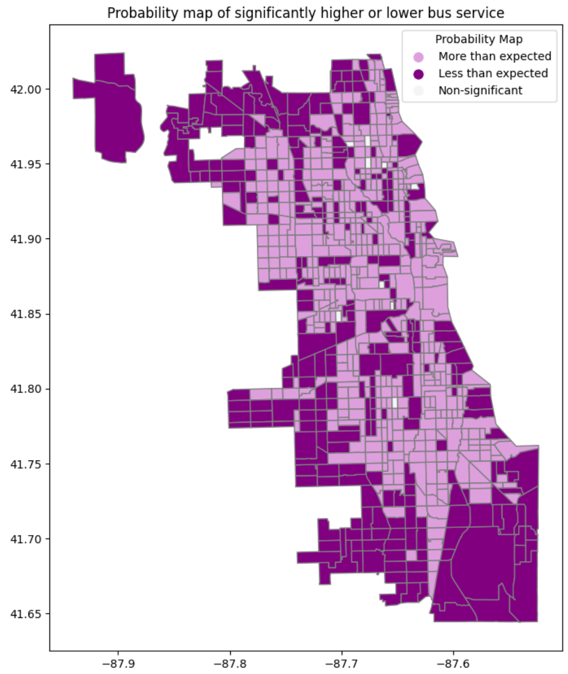
Again we see higher expected service rates for larger census tracts by area and in the actual plot we see that the loop and areas around the economic center have higher counts of stops than expected. Some slightly lighter areas around Hyde Park and in the Loop surroundings when compared to their expectation indicate that these areas might have more service than expected. And slightly darker areas in the outskirts indicate that these might be slightly under serviced areas.

To compare the significance of how much these rates varied from their expectation, I first used a poisson distribution to determine probabilities. Given that census tracts ranged from having 0 bus service stops in their tract (5 tracts that are likely outside the range of CTA service), up to 28,934 in the West Loop there was a vast range of service counts, and tracts varied widely. As we see in the poisson distribution, poisson probabilities were very low indicating large variation from the expection.

Probability Map of Bus Stop Counts Across Census Tracts Using Poisson Distributions

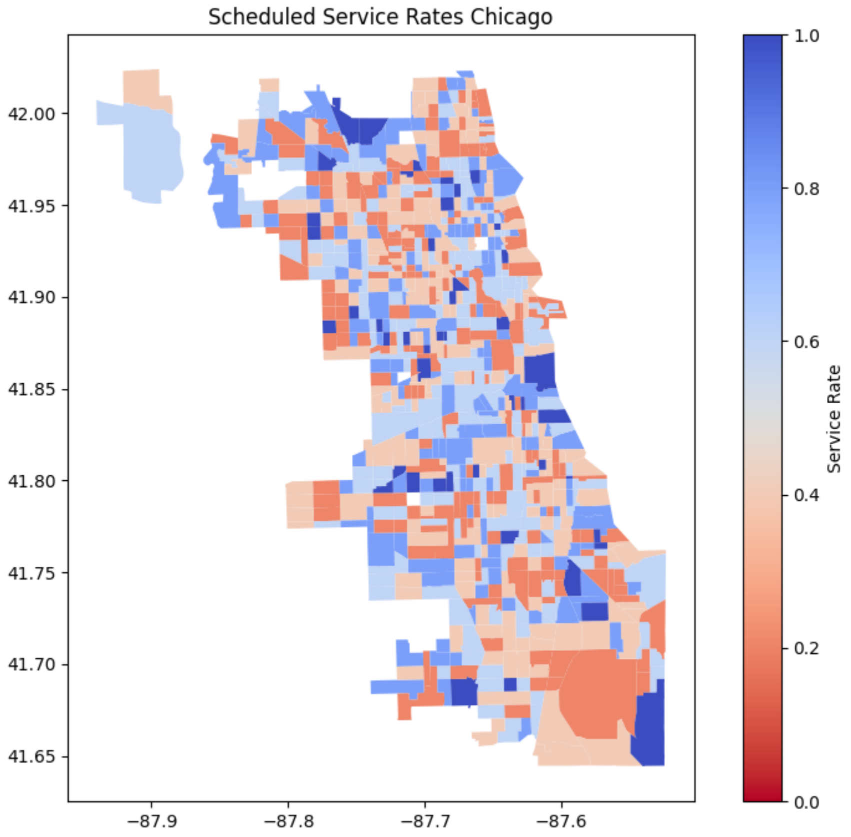


To parse apart the directionality of the variation (if tracts had lower or higher levels of service), I plotted according to their count and if they had a statistically significant probability.



As would probably be expected, more central areas had more stops than expected, but there are a few surprising gaps in the south side that seem to be underserviced.

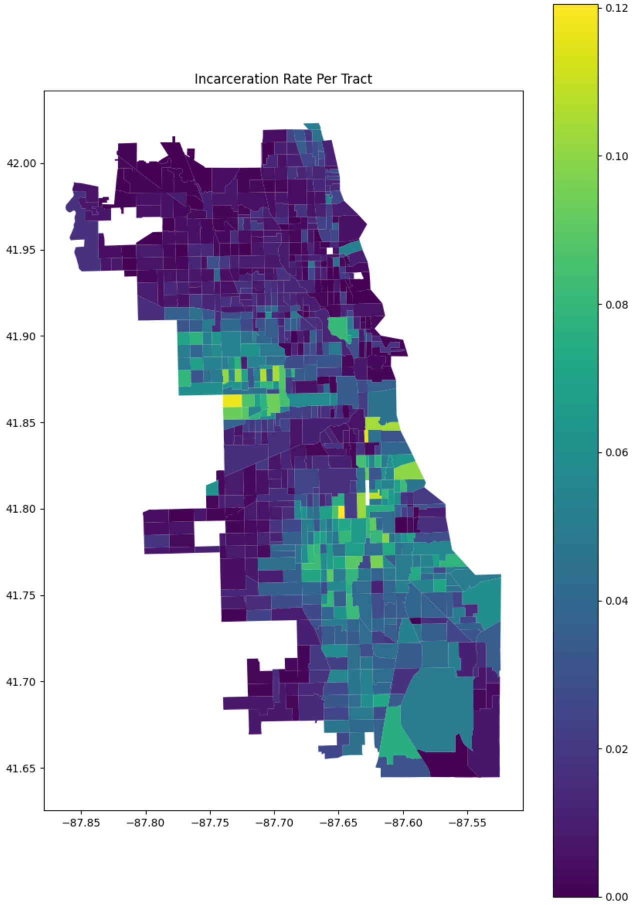
Finally to get a better sense of the spectrum of scheduled service I used hinge15 breaks to bucket the service rates according to the distribution and get a more nuanced view of the extent to which areas had more or less service rather than purely above or below and statistically significant.

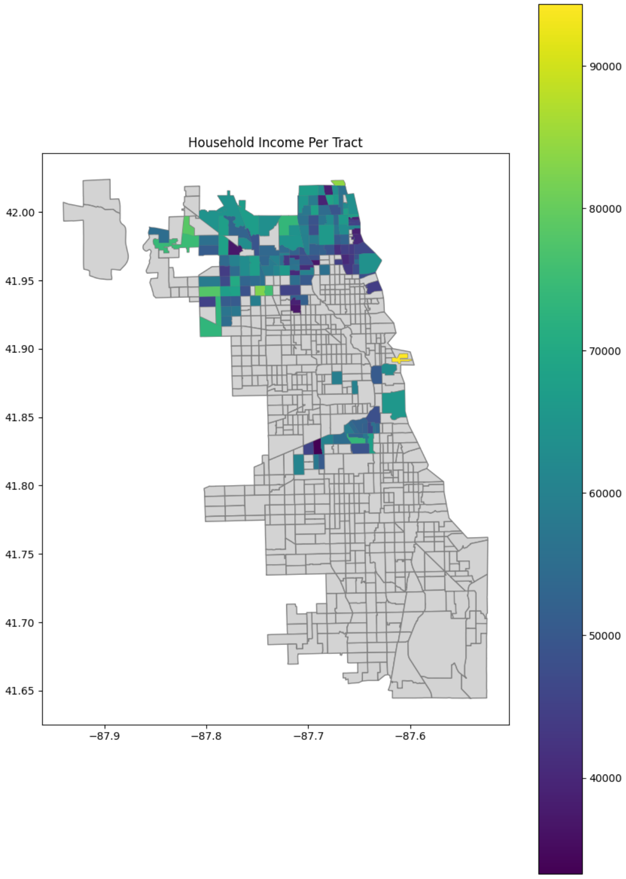


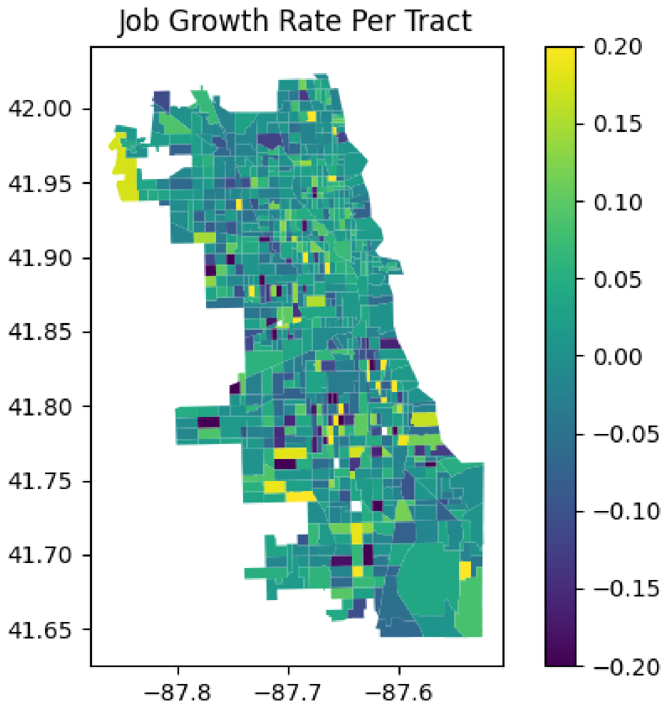
From this we can see again that the loop is particularly over-serviced which makes sense given the prioritization of economic centers in a city. And the "underservicing" seems to be rather evenly distributed between North and South in what seems to be the more residential and single-family neighborhoods of Chicago.

3.4 Socioeconomic Indicator Analysis

To gain a sense of whether these transit patterns are signals for other trends of inequity in Chicago, I wanted to compare against socioeconomic factors. Census tract-level data was gathered for job growth, household income and incarceration rates. There was substantial missingness in household income and education was not available at a tract level of granularity, otherwise those would have been informative.







From these plots of each socioeconomic factor by census tract, we see that household income is not sufficiently representative, job growth has small fluctuations but no clear trends, and incarceration has the strongest pattern shown by the lighter South and central West patterns (this may also be the result of census averaging).

Despite the unlikelihood of these patterns having any correlation with bus service trends from visual inspection, I ran a simple linear regression for scheduled bus service on each census characteristic. The results found no significant relationship for job growth or household income, however it did find a slightly statistically significant relationship between incarceration rate and service, with a p-value of 0.024 for a coefficient of $8.168e^{-}7$ indicating that it is an extremely small (nearly 0) relationship, but still worth noting.

4 Discussion

Given the aforementioned methods and exploration, we can now evaluate the initial hypotheses.

1. For hypothesis 1, we **reject the null hypothesis** in favor of the alternative given that the probability map of bus stops per census tract shows statistically significant differences in many tracts.
2. For hypothesis 2, we again **reject the null hypothesis** given that the probability map of scheduled service again shows statistically significant differences in service between census tracts. Over and under-servicing seemed to be more extreme when considering scheduled service than in raw bus stop counts.
3. For hypothesis 3, **the null hypothesis holds** as we did not see any significant correlation between census opportunity indicators and bus service. This may be worth further investigation on different levels of granularity, however given the metrics for this scope, nothing worthwhile was found.

From this analysis it is clear that there is variation of transit service throughout Chicago and access is particularly centered around the loop and the North side. Most areas that lack access are in the South and central West parts of Chicago, which also overlap with many historical and current socioeconomic trends of disadvantage.

5 Conclusion

This analysis explored a few different approaches to assessing transit equity in Chicago. By assessing bus stop location point data, we could better understand point density patterns and the likelihood of access to nearby transit stops. We found that bus stop access is easy along core economic areas and big streets (Milwaukee, the Loop, King Drive), but this access does not carry particularly into primarily residential areas.

In exploring densities we can determine that in aggregate, some census tracts lack bus stop access, while others have more than the expectation. These patterns fall along predictable far North West and Central West, South areas lacking access, and North and central areas having more access. Assessing scheduled service essentially amplified the trends we saw from bus stop access.

Comparing these patterns of access against socioeconomic indicators did not show any statistically significant trends, however there are likely many other metrics to consider in this context.

Further analysis could benefit from incorporating two main factors. The first would be ghost bus data scraped from live transit feeds on a regular basis in order to get a more realistic sense of actual service. Since CTA has a history of not providing service at the rate that schedules indicate, it would be important to asses how the actual service varies throughout Chicago. From this we can assess if it further amplifies the trends we saw, or if poor service is equally allocated. Additionally, it would be beneficial to consider a range of other socioeconomic indicators that may be directly related to transit access such as walkability, cars owned per individual, average commute time, and method of commuting. These factors would shed light on the demand a given community has for transit, and the impact that improved transit access may have on them.

In summary, we now see that access to transit varies widely across Chicago affecting some Chicagoans more than others, thereby limiting their access to resources and opportunity. By visualizing these discrepancies, hopefully this can help to inform paths to improving access for all.

6 Code

The Github repository where all data analysis was conducted can be found here (if the github hyperlink does not work: https://github.com/meganhmoore/TransitClustering)