Relationship Between Public Transport Accessibility and Car Dependency in Chicago[1]

GGR473 - By: Jay, Sabih, Missy

Introduction

Transportation plays a vital role in accessing employment, education, healthcare, and other essential services. However, low-income households often face significant challenges in accessing reliable and affordable transportation options. This issue has been the focus of numerous studies, highlighting the need to address the transportation challenges faced by low-income households. Public transit offers an affordable and safe way of getting around without requiring to purchase and maintain a vehicle. In this context, our group project aims to explore the relationship between proximity to public transport, car ownership, and the income of surrounding residents. This report will provide background information on the importance of addressing this challenge, outline our research methods, and explore our key findings and results.

Project Background:

Various studies have highlighted the challenges faced by low-income households regarding transportation options. Owning a car enables better access to employment opportunities and higher incomes for individuals in car-dominant societies (Blumenberg & Thomas, 2014). However, car ownership among low-income communities can be unstable and subject to fluctuations influenced by life events, such as job loss or deteriorating health conditions (Klein & Smart, 2019). This poses a significant burden on these households' financial stability, potentially trapping them in a cycle of poverty (Curl et al., 2018).

[1] Data source and code are available at: https://github.com/uoft-flyfreejay/ggr473-final-project

Simultaneously, evidence suggests that improvements in transit accessibility have the potential to shift mode choice in low-income neighborhoods. According to studies by Ellen and O'Regan (2011) and Pucher and Renne (2003), if owning a private vehicle poses a significant burden, low-income households may be more responsive to transit accessibility improvements and more likely to switch from cars to public transit. Therefore, understanding the relationship between proximity to public transport and car ownership is crucial. It can provide insights into whether individuals choose to live near public transport out of financial necessity or if proximity to public transport reduces car dependency.

Research Questions:

Our group project aims to explore the relationship between proximity to public transport, car ownership, and the income of surrounding residents. The research questions we seek to answer are:

- Does the density of bus stops in one's neighborhood affect car dependency?
- Is there a correlation between income and car dependency?
- Does proximity to public transport reduce car dependency among low-income households?

The exploration of these research questions is essential for urban development. Firstly, it has important policy implications. Understanding the relationship between public transport access and car ownership can provide valuable insights for policymakers and urban planners. This knowledge can assist in designing transportation systems that cater to the needs of low-income households. If proximity to public transport significantly reduces car dependency, it emphasizes the need for enhanced accessibility and improved public transit options for these individuals. Furthermore, the research also aligns with sustainable urban development goals. Encouraging a shift from car ownership to public transport is crucial for promoting sustainable transportation options. Understanding the factors that influence mode choice, such as proximity to public transport, can aid in developing effective strategies to reduce car dependency.

By exploring these research questions, we hope to contribute valuable insights to inform policymakers, urban planners, and community organizations in developing strategies and interventions that enhance public transport accessibility and reduce car dependency among low-income communities.

Methods

We obtained our data from two sources, the Chicago Data Portal and U.S. Census Bureau.

Our desired data for each census tract is as follows:

- Median income
- Density of bus stops per hectare
- Nearest distance to bus stop from census centroid (m)
- Percentage of households with at least 1 car

Data sources we used:

[1] Household median income

[2] Vehicles per household

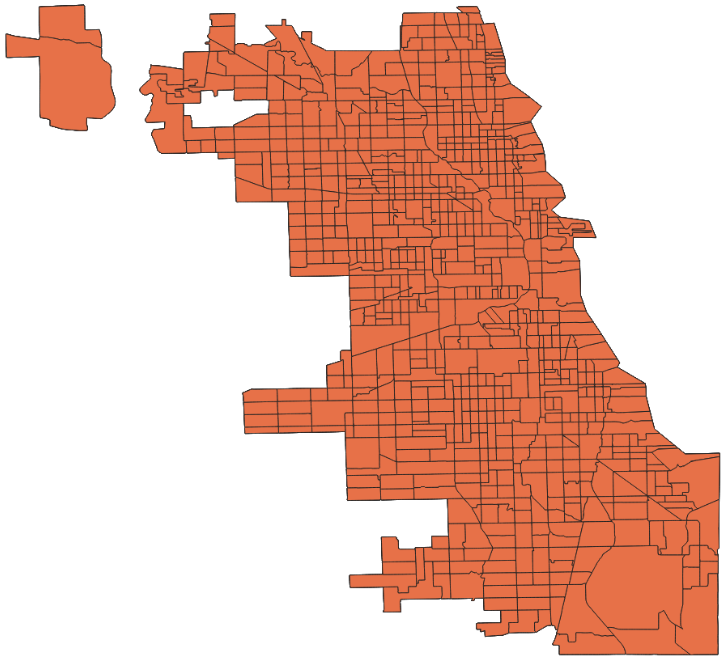
[3] Census tract boundaries

[4] Bus stops in Chicago

[1] Household median income was given as a .csv. It had the 1 column we wanted, median income.

[2] Vehicles per household was given as a .csv. It did not have the final column we wanted, the percentage of households with at least 1 car. We extracted 4 columns, which were households with either 0, 1, 2, or 3+ vehicles, to calculate the final percentage.

[3] The census tract boundary was given in GeoJSON format. The original boundaries look like below when loaded into QGIS:



[4] Bus stops were also given in GeoJSON format as a point layer.

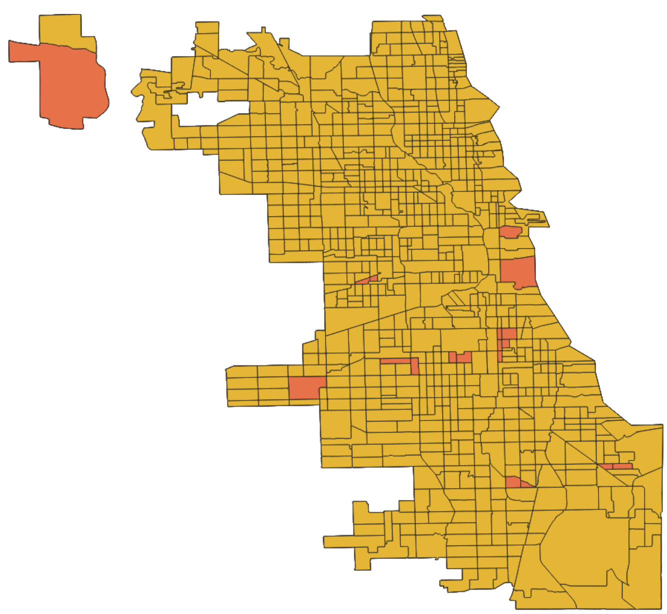
The projection of all the data layers was reprojected to EPSG:2958.

To create the database from the raw files, following steps were followed.

1. Utilized Excel to extract certain columns that contain information on vehicle ownership and income as well as rename them to more intuitive column titles. Save it as a new census file. This now includes columns that represent census tract number, number of houses with no vehicles, one vehicle, two vehicles, and three or more vehicles, and median income.
2. Import the boundary file as a geospatial table and the extracted census data onto a created database using QGIS database manager.

3. Perform data cleaning on census data using SQL queries. It involves:
    a. Keeping the common key consistent, such as editing Census Tract value to not have trailing address after the track number
    b. Convert strings to numbers, as well as omit non-numeric characters like "+" from the value
    c. Drop rows with null values. They indicate missing census or invalid data
4. Insert a new column with the percentage of vehicle ownership. To do this, we added all values in columns that represent the number of households with one, two, three or more vehicles and divided it by the sum of all households per census tract. Multiply the result by 100% to get the percentage.
5. Join the geospatial table that contains boundary data and the table with processed census data using the primary key that represents the census tract number.

At the end, our final database involves columns with numerical values that each represent the number of households that own no vehicle, one vehicle, two vehicles, three or more vehicles, median income, vehicle ownership rate, the census tract number, as well as other geospatial data. The following view is created at the end, note that the orange regions have been removed due to a prior step that removed census tracts with null values for vehicle data:

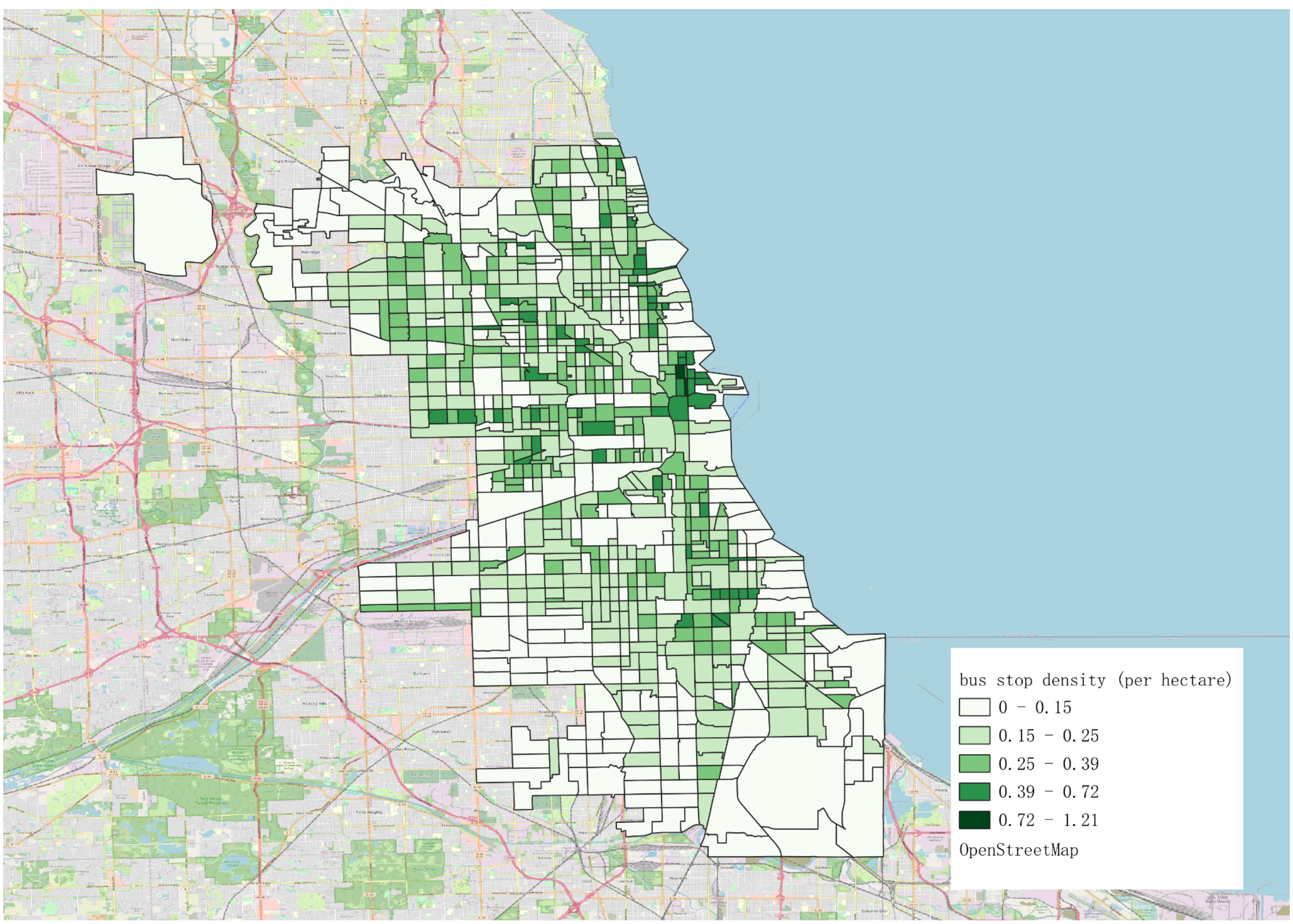


Bus stop density per census tract:

Next we wanted to find the bus stop densities. To do this we used the census tract boundary layer as well as the bus stops point layer. These steps were taken:

1. Use the field calculator with $area operation to measure the polygon area for each census tract. Add this as a new column called area to the census tract layer.
2. Use the "count points in the polygon tool" to count bus stops in each census tract. Add these counts as a new column called points to the census tract layer.
3. Divide the area column by the number of points column to get the bus stop density per census tract. Add this as a new column to the census tract layer.
4. We also added this column to the "code/data_for_linear_reg.csv" file.

To create the map of bus stop density, we then went into the layer properties of the census tract, and symbolized the density field. We used natural breaks with 5 classes. Here is the resulting map:



Nearest bus stop distance to every census centroid:

To get the nearest bus stop distance from the centroid of each census tract, we followed these steps:

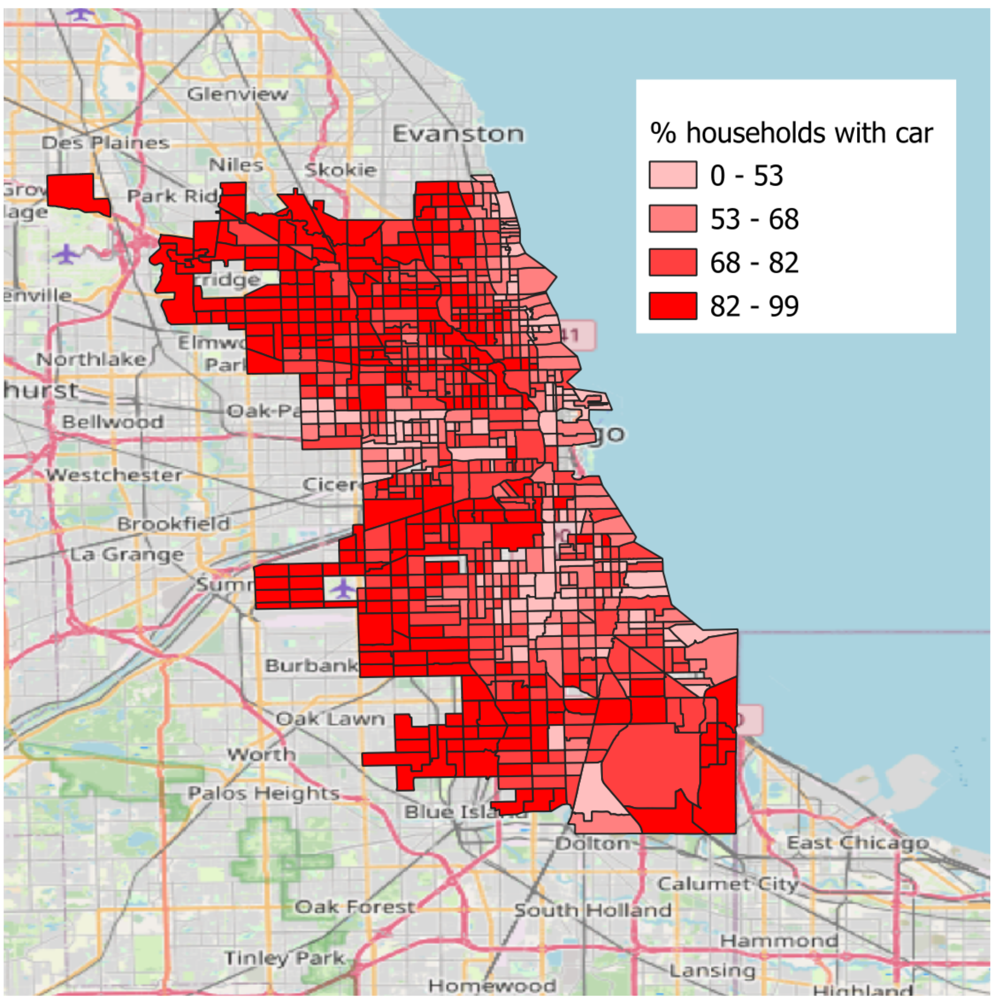
1. Use the centroids tool in QGIS to make a point layer of each census tract boundary
2. We now have a point layer of the census centroids and a point layer of the bus stops. We use the "Distance to nearest hub (points)" tool with the census centroids as the source layer and bus stops as the destination layer. For every point in the census centroid, it will find the distance to the closest bus stop. This value is stored in a new column in the census centroids layer.
3. We also added this column to the "code/data_for_linear_reg.csv" file.

Percentage of households with a car per census tract:

To find the % of households with at least 1 car, we used these steps:

1. From the "Vehicles per Household" data, we ran the python script under "code/calculate_percent_households_with_car.py". It summed up the number of households with at least 1 car over the total number of households, times 100.
2. We also added this column to the "code/data_for_linear_reg.csv" file.

Here is a QGIS map of the data:



Results

The results will be composed of 2 parts: the linear regression done on multiple and singular variables, as well as an analysis of the bus stop density map alongside the car ownership map.

Let's begin by looking at the data we have in "code/data_for_linear_reg.csv" file." We will perform a multiple linear regression to see the effect on the percentage of households with a car. The code we will run is all stored in "code/linear_regression.py". An R-squared value falls from 0 to 1. A value of 0.5 means 50% of the effect on the dependent variable correlates with the independent variable.

Variables: Median income, Bus stop density, and Nearest distance to bus stop

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     percent_households_with_car   R-squared:                       0.189
Model:                                     OLS   Adj. R-squared:                  0.186
Method:                          Least Squares   F-statistic:                     59.92
Date:                         Wed, 29 Nov 2023   Prob (F-statistic):           8.12e-35
Time:                                 22:07:55   Log-Likelihood:                -3129.2
No. Observations:                          773   AIC:                             6266.
Df Residuals:                              769   BIC:                             6285.
Df Model:                                    3
Covariance Type:                     nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      75.6725      1.549     48.842      0.000      72.631      78.714
density(/hectare)         -42.0159      4.424     -9.498      0.000     -50.700     -33.332
Median income (dollars)     0.0001   1.21e-05      8.362      0.000    7.71e-05       0.000
Nearest distance (km)       0.0020      0.002      1.033      0.302      -0.002       0.006
==============================================================================
Omnibus:                       37.399   Durbin-Watson:                   0.347
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.119
Skew:                          -0.550   Prob(JB):                     1.18e-09
Kurtosis:                       2.745   Cond. No.                     7.51e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.51e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In the top right the R-squared value is 0.189. Quite weak. Let's try the other combinations.

Variables: Bus stop density and median income

```
                              OLS Regression Results
==============================================================================
Dep. Variable:     percent_households_with_car   R-squared:                       0.188
Model:                            OLS   Adj. R-squared:                  0.186
Method:                 Least Squares   F-statistic:                     89.34
Date:                Wed, 29 Nov 2023   Prob (F-statistic):           1.29e-35
Time:                        22:05:06   Log-Likelihood:                -3129.8
No. Observations:                 773   AIC:                             6266.
Df Residuals:                     770   BIC:                             6280.
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    76.5107      1.320     57.961      0.000      73.919      79.102
density(/hectare)       -43.6818      4.119    -10.604      0.000     -51.768     -35.595
Median income (dollars)   0.0001    1.2e-05      8.410      0.000    7.76e-05       0.000
==============================================================================
Omnibus:                       37.396   Durbin-Watson:                   0.344
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.229
Skew:                          -0.552   Prob(JB):                     1.11e-09
Kurtosis:                       2.754   Cond. No.                     6.93e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.93e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Once again quite weak at 0.188 but pretty much the same as before.

Variables: Nearest bus stop and Median income

```
                              OLS Regression Results
==============================================================================
Dep. Variable:     percent_households_with_car   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                  0.092
Method:                 Least Squares   F-statistic:                     40.12
Date:                Wed, 29 Nov 2023   Prob (F-statistic):           2.66e-17
Time:                        22:08:31   Log-Likelihood:                -3172.1
No. Observations:                 773   AIC:                             6350.
Df Residuals:                     770   BIC:                             6364.
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    65.1300      1.142     57.039      0.000      62.888      67.371
Median income (dollars) 9.635e-05  1.27e-05      7.574      0.000    7.14e-05       0.000
Nearest distance (km)     0.0089      0.002      4.570      0.000       0.005       0.013
==============================================================================
Omnibus:                       43.083   Durbin-Watson:                   0.177
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.610
Skew:                          -0.564   Prob(JB):                     1.25e-10
Kurtosis:                       2.622   Cond. No.                     1.78e+05
==============================================================================

Notes:
```

The weakest we've seen so far at 0.094. Half as weak as the previous 2 variables.

Variables: Bus stop density

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     percent_households_with_car   R-squared:                 0.114
Model:                                     OLS   Adj. R-squared:            0.113
Method:                          Least Squares   F-statistic:               98.99
Date:                         Wed, 29 Nov 2023   Prob (F-statistic):     5.04e-22
Time:                                 22:00:14   Log-Likelihood:          -3163.7
No. Observations:                          773   AIC:                       6331.
Df Residuals:                              771   BIC:                       6341.
Df Model:                                    1
Covariance Type:                     nonrobust
=====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               83.4929      1.072     77.909      0.000      81.389      85.597
density(/hectare)  -42.7848      4.300     -9.949      0.000     -51.226     -34.343
==============================================================================
Omnibus:                       46.237   Durbin-Watson:                   0.214
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.498
Skew:                          -0.639   Prob(JB):                     2.42e-12
Kurtosis:                       2.841   Cond. No.                         8.63
==============================================================================
```

0.114, quite weak as well.

Variables: Median income

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     percent_households_with_car   R-squared:                 0.070
Model:                                     OLS   Adj. R-squared:            0.069
Method:                          Least Squares   F-statistic:               57.87
Date:                         Wed, 29 Nov 2023   Prob (F-statistic):     8.18e-14
Time:                                 22:02:59   Log-Likelihood:          -3182.5
No. Observations:                          773   AIC:                       6369.
Df Residuals:                              771   BIC:                       6378.
Df Model:                                    1
Covariance Type:                     nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                      67.2380      1.058     63.558      0.000      65.161      69.315
Median income (dollars)  9.797e-05   1.29e-05      7.607      0.000    7.27e-05       0.000
==============================================================================
Omnibus:                       44.600   Durbin-Watson:                   0.123
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               46.902
Skew:                          -0.570   Prob(JB):                     6.54e-11
Kurtosis:                       2.604   Cond. No.                     1.62e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+05. This might indicate that there are
```

Quite weak as well at 0.07.

Variables: Nearest bus stop

```
                             OLS Regression Results
=========================================================================================
Dep. Variable:     percent_households_with_car   R-squared:                       0.027
Model:                                     OLS   Adj. R-squared:                  0.026
Method:                          Least Squares   F-statistic:                     21.32
Date:                         Wed, 29 Nov 2023   Prob (F-statistic):           4.56e-06
Time:                                 22:02:01   Log-Likelihood:                -3199.9
No. Observations:                          773   AIC:                             6404.
Df Residuals:                              771   BIC:                             6413.
Df Model:                                    1
Covariance Type:                     nonrobust
===========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                   71.8554      0.744     96.626      0.000      70.396      73.315
Nearest distance (km)    0.0093      0.002      4.617      0.000       0.005       0.013
==============================================================================
Omnibus:                       47.585   Durbin-Watson:                   0.056
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.760
Skew:                          -0.585   Prob(JB):                     1.57e-11
Kurtosis:                       2.580   Cond. No.                         504.
==============================================================================
```

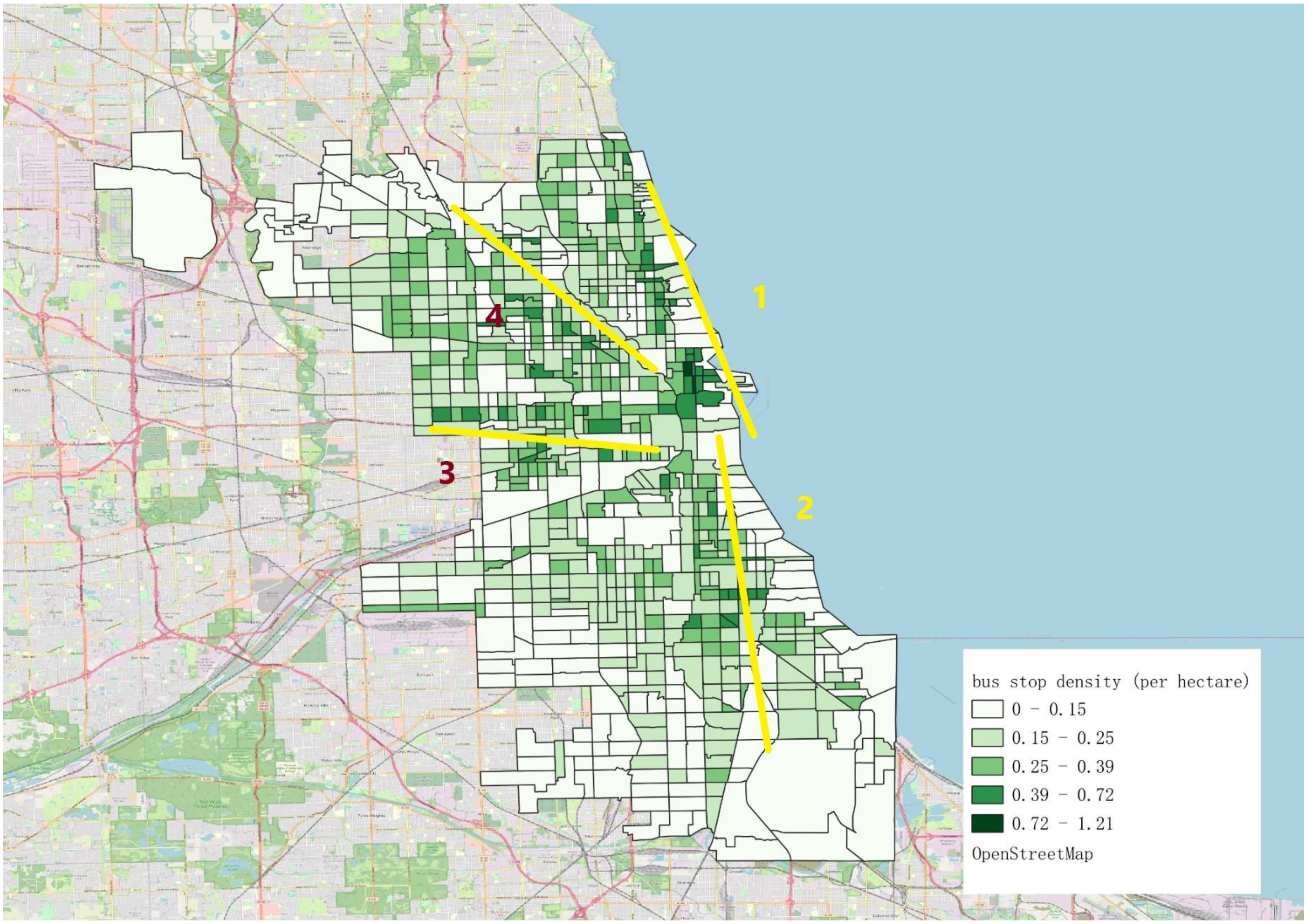
Weakest correlation so far at 0.027.

The highest R-squared is obtained with all 3 variables, or arguably, the same as just bus stop density + median income, rounded up to 0.19. Bus stop density has the highest singular R-squared with 0.114.

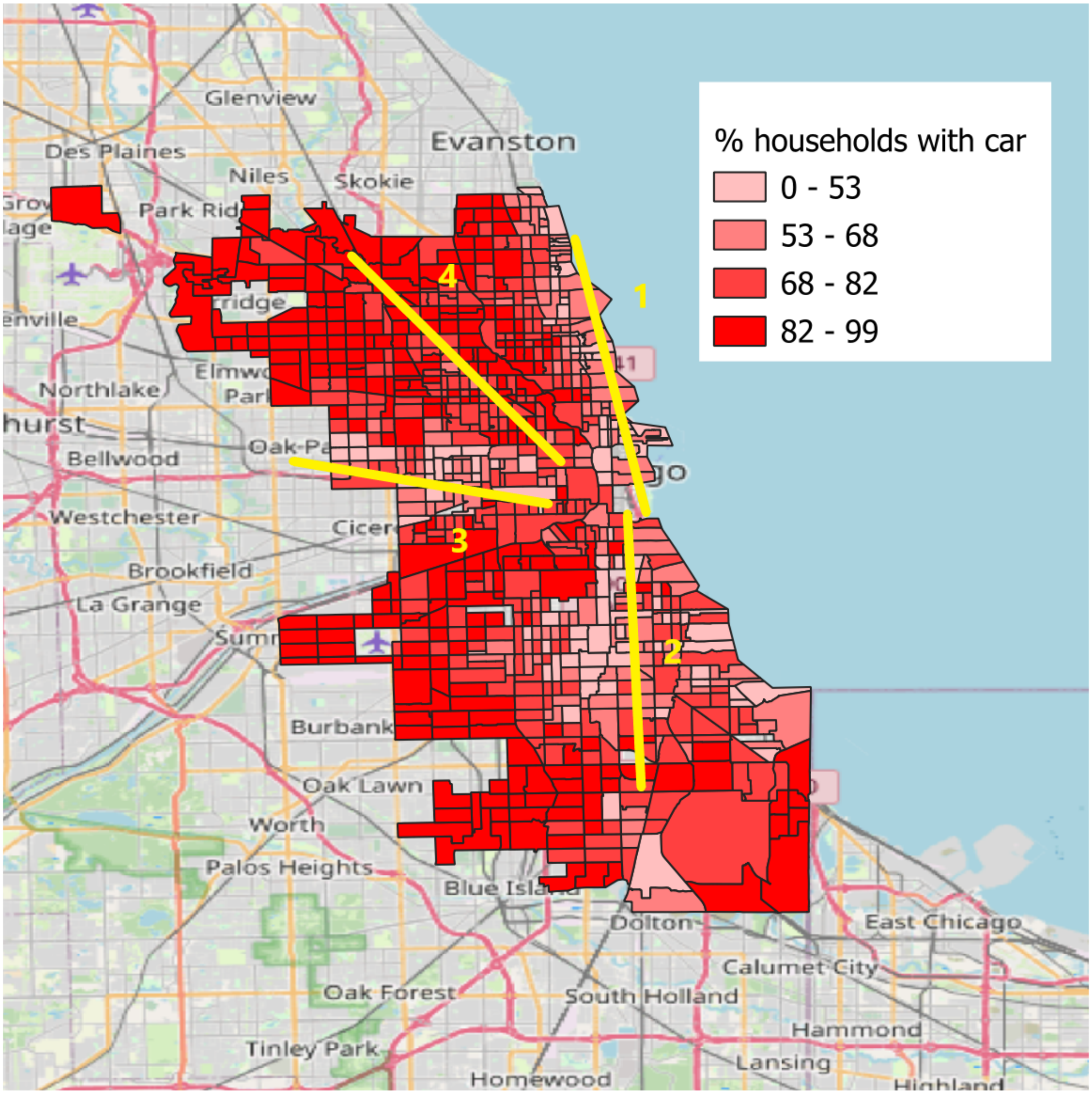
It is interesting how of the combinations of the 2 variables, bus stop density + median income is almost the same as all 3 variables. For this reason, we believe the nearest distance to a bus stop from the centroid is the weakest variable. The strongest predictor is bus stop density at 0.114, followed by median income at 0.07. They are relatively weak though.

Let's take a look at the bus stop density maps and visually try to find a pattern. I have added 4 colored lines as reference in each.

Bus density annotated map:



Households with a car annotated map:



In both maps I drew the same lines and labeled them from 1 to 4. Notice that in the density map, near the 4 lines you see there is a corresponding line of high bus density indicated by green. If you look at the same lines on the households with a car, lines 1 to 3 have households with lower car ownership. Line 4 has high bus density but also high car dependency for some reason. I don't know why but a highway does pass through line 4 which may encourage residents to have a car to travel further.

Based on these maps, I have a hunch that we see a noticeable decrease in car ownership as bus stop density reaches a critical threshold, roughly 0.4 stops per hectare. Further analysis would be needed to test this theory.

In the future we also want to look at population density and building density to see how they are correlated with car ownership and bus stop density.

Discussion

The results of our study suggest that there is a weak relationship between public transport accessibility and median income on car dependency in Chicago. We analyzed multiple linear regression models using variables including median income, bus stop density, and nearest distance to a bus stop. In all cases, the R-squared values were relatively low, which suggests that the variables we examined have a limited impact on car ownership within the city.

While Blumenberg and Thomas' research underscores the challenges faced by low-income communities and their reliance on public transit (Blumenberg & Thomas, 2014) , our project unveils different dimensions of this complex issue. Our results suggest that proximity to public transport alone may not significantly reduce car dependency among low-income households in Chicago. Other factors such as cultural norms and infrastructure design may play a more significant role in determining car ownership patterns. This indicates that the relationship between accessibility of public transport and car dependency in low-income households is nuanced and influenced by various factors.

The development of the spatial database used in our research is valuable for future studies in several ways because it provides a foundation for analyzing the correlation between public transport accessibility, car dependency, and other socio-economic factors. Researchers can build upon this database to conduct further investigations and explore additional variables that may influence transportation patterns. However, additional data are needed to enhance the understanding of the relationship between public transport accessibility and car dependency. Travel behavior data can provide insights into commuting patterns and travel mode choices to help explore the decision-making process behind car ownership. Land use data can provide information on the distribution of residential, commercial, and mixed-use developments, which can influence transportation choices. More socio-economic data such as population density and employment rates can provide further context for determining car ownership patterns. By incorporating these additional data sources, future research can deepen the understanding of the relationship between public transport accessibility and car dependency.

Critical Reflection Piece

Challenges

One challenge we faced was finding the vehicles per household data from the US Census portal. There is no dataset that was specifically about this. One did exist called “Commuting characteristics by sex” but this did not actually contain the number of vehicles per household. We had to find it in the dataset named “Physical Housing Characteristics”. It had 4 columns indicating the number of households with 0, 1, 2, or 3+ vehicles. By randomly opening datasets we come upon the data we wanted.

Another challenge was that the field calculator in QGIS was buggy. We wanted to use the 4 vehicle columns mentioned above to calculate the percentage of households with a car. But the calculated output column was null for some reason. QGIS was unable to read the 3 input columns properly. To overcome this, I used a python script that did the same thing.

When working with GIS software like QGIS for calculating distances, such as finding the nearest bus stops, it's essential to a consistent coordinate system. For our project, we set the EPSG code to 2958. This won't permanently change the map layers, but it'll temporarily transform them for accurate measurements in kilometers. With a projected Coordinate System (PCS) like EPSG 2958, distance is shown in kilometers.

Limitations

The primary limitation encountered was the inherent constraints of census data. Census data, while comprehensive, often lacks granularity. Data is aggregated at the census tract level which has arbitrary borders. We are not able to create a grid of equally shaped hexagons instead for spatial analysis.

Additionally, the Chicago Census Boundary Data was quite outdated. We could only find the 2010 census tracts so all of our data is from 2010 as well. The exception is the bus stop data which is from 2013. Getting the same year data other than the US Census was difficult. Only the US Census has the resources and planning to regularly update data.

Another significant limitation was the use of census tract centroids to calculate distances to bus stops. This method oversimplifies the geographical diversity within each tract and can lead to inaccurate estimations of public transport accessibility. The centroid of a census tract might not accurately represent the actual living area of its residents, especially in tracts with irregular shapes or varied topography.

Furthermore, as bus stops were the only public transit we used for our analysis, we did not account for other methods of transportation such as subways. They could have provided further insights as well as stronger correlation. Within the bus stops, we could have also looked for other characteristics such as bus frequency per station, its accessibility, the capacity, and etc. However, looking for such detailed information per bus stop would have also required researching for other resources and datasets as they are very specific and some of the specific details may not even be accessible unless one directly goes in person for field surveying.

Ethical Considerations

From an ethical standpoint, the use of the data posed questions about representation. When using neighborhood-level measures for correlation analysis, one ethical concern is the risk of ecological fallacy, which is the assumption that characteristics at the aggregate level can be directly applied to individuals within that group, leading to potential inaccuracies and misinterpretations.

In our study, correlations between proximity to public transit, income levels, and car dependency at the neighborhood level cannot be automatically assumed that these relationships hold true for every individual within that neighborhood. Individuals may not conform to the average characteristics of their neighborhood and that specific circumstances and variations exist within any given group. Failing to account for these variations can lead to unintended consequences and reinforce existing inequalities or biases.

Implications for Further Analysis

If the current data definitions and methodologies were to be used for additional analysis the lack of consideration for other forms of public transportation, such as the subway or bikes, limits the understanding of urban mobility and accessibility. In addition, the data we used in our analysis provide correlations between variables, but they do not establish causal relationships. We cannot definitely conclude that changes in one variable will directly cause changes in another.

In summary, based on our project, future research would benefit from more granular, up-to-date data, a more nuanced approach to spatial analysis, and a broader consideration of urban characteristics to more accurately and ethically represent the picture of car ownership in Chicago.

References:

Data Sources:

U.S. Census Bureau. (2010). MEDIAN INCOME IN THE PAST 12 MONTHS (IN 2010 INFLATION-ADJUSTED DOLLARS) . *American Community Survey, ACS 5-Year Estimates Subject Tables, Table S1903*. Retrieved December 6, 2023, from https://data.census.gov/table/ACSST5Y2010.S1903?q=vehicle&t=Income (Households, Families, Individuals)&g=040XX00US17$1400000&y=2010.

U.S. Census Bureau. (2010). PHYSICAL HOUSING CHARACTERISTICS FOR OCCUPIED HOUSING UNITS . *American Community Survey, ACS 5-Year Estimates Subject Tables, Table S2504*. Retrieved December 6, 2023, from https://data.census.gov/table/ACSST5Y2010.S2504?q=vehicle&g=040XX00US17$1400000&y=2010.

Chicago Data Portal (2013). Boundaries - Census Tracts - 2010. Retrieved December 6, 2023, from https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik

Chicago Data Portal (2017). CTA Bus Stops. Retrieved December 6, 2023, from https://data.cityofchicago.org/Transportation/CTA-Bus-Stops/hvnx-qtky

Journals:

Blumenberg, E., & Thomas, T. (2014). Travel Behavior of the Poor after Welfare Reform. *Transportation Research Record*, *2452*(1), 53–61. https://doi.org/10.3141/2452-07

Curl, A., Clark, J., & Kearns, A. (2018). Household car adoption and financial distress in deprived urban communities: A case of forced car ownership? *Transport Policy*, *65*, 61–71. https://doi.org/10.1016/j.tranpol.2017.01.002

Ellen, I. G., & O’Regan, K. M. (2011). How low income neighborhoods change: Entry, exit, and enhancement. *Regional Science and Urban Economics*, *41*(2), 89–97. https://doi.org/10.1016/j.regsciurbeco.2010.12.005

Klein, N. J., & Smart, M. J. (2019). Life events, poverty, and car ownership in the United States: A mobility biography approach. *Journal of Transport and Land Use*, *12*(1), Article 1. https://doi.org/10.5198/jtlu.2019.1482

Pucher, J., & Renne, J. L. (2003). *Socioeconomics of urban travel. Evidence from the 2001 NHTS.* https://orlis.difu.de/handle/difu/126088

Github Repository:

FlyFreeJay. (2023). ggr473-final-project. GitHub. https://github.com/uoft-flyfreejay/ggr473-final-project