Exploring Social Disparities across Neighborhoods in Chicago

Junjun Zheng

Jun 2021

INTRODUCTION

Urban areas would accommodate over 70% world population by 2050. Modern cities are organized to effectively build out civic infrastructure and efficiently deliver necessary public services. However, there are growing concerns of social inequality that urban residents may have inequitable access to infrastructure and services depending on where they live.

Social injustice results in imbalanced economic development and contributes to gaps in quality of life across urban neighborhoods. Yet, despite increasing awareness toward social inequity, quantitative assessments for such social disparities are still lagging. This situation becomes the barrier to sustainable development, and inadequate understanding of the underlying conditions hinders strategic urban planning for potential implications.

This study deploys data science toolkits to explore social disparities across neighborhoods. The study develops a novel framework by integrating diverse analytic approaches. The numerical analysis quantifies the relationship among various neighborhood characteristics, including household income, spatial clusters, and land uses. Therefore, this study bridges the neighborhood's socioeconomic factors with the infrastructure and built environments, aiming to provide nuances to equitable community investment and place-based development initiatives.

The following section describes the data and methods for the study, involving data collection, cleaning, and analytical methods. The results present both the descriptive and exploratory analyses, as well as data visualization. This study also concludes the take-aways from the numerical analysis and discusses potential implications for the community and local stakeholders.

DATA AND METHODS

The numerical analysis takes the city of Chicago as the case study area. Chicago consists of 56 zip codes (neighborhoods), which is the unit of analysis in this study. Based on the existing projects and studies, the zip code is one of the best scales to analyze the spatial heterogeneities within a city.

Data Acquisition and Cleaning

By looking at zip codes, this study collects data from diverse sources. The base map consists of the polygons of all Chicago zip codes in a *shapefile*[1]. On the one hand, *median household income* (MHI of each zip code) is the proxy for socioeconomic factors. U.S. Census provides the data under American Community Survey (2019, 5-year estimates).

On the other hand, *Foursquare location data* are collected for quantifying neighborhood land uses. First, all Chicago zip codes are indexed to their centroid based on the latitudes and longitudes of the polygons. Then, this study collects all venues within 800 meters of each neighborhood centroid by using *Foursquare API*, which results in 5,375 venues that meet the requirement.

However, these over 5000 venues have 423 unique categories, which are too detailed to categorize neighborhood land-use patterns. For example, the restaurants and food-service outlets have been categorized into over 30 categories: American, Chinese, Japanese, etc. A data cleaning process is needed for further analysis. Hence, this study pulls out over 400 unique categories and crosswalks them to ten (10) major categories, including:

- Offices,
- Professional Services,
- Retails,
- Restaurants,
- Schools,
- Sport and fitness,
- Leisure,
- Civic Services,
- Transportation, and
- Other Land Uses

The crosswalk data (provided in data files) reduces the dimensions of the original dataset and helps improve the efficiency and accuracy of the analysis.

Furthermore, over 5000 venues are aggregated into 56 zip codes by the ten business categories. This calculates the shares of each business category against total businesses for each zip code. For example, the "*Offices"* under the zip code *"60610"* is 0.5163, indicating 51.63% of businesses within this zip code are offices. As a result, the data consist of land use patterns for all Chicago zip codes.

Once all data are collected and cleaned, this study integrates a dataset by merging the socioeconomic variable with the land use data. The final dataset includes variables of geometries, median household income (MHI), and land use shares (for ten major categories).

[1] The shapefile format is a geospatial vector data format for geographic information system software.

Methodology

This study deploys the k-means clustering analysis to identify the typologies of Chicago neighborhoods. This approach aims to partition *n* observations into *k* clusters based on input variables. All observations belonging to the same cluster have the nearest mean distances (to the cluster centroid). Numerically, k-means clustering minimizes within-cluster variances and optimizes squared errors. At the same time, the key challenge for implementing k-means clustering is to determine the optimal *k* – the number of clusters. This study, therefore, computes *silhouette scores* for each *k* option to find the best partitioning result.

As described in the previous section, the complete dataset consists of 13 variables for clustering analysis (***Table 1***). However, their ranges and variances differ greatly, and these variables need to be standardized before implementing the clustering algorithm. Also, this study involves geometric variables (i.e., latitude and longitude) for clustering. This is because the spatial spillover appears in urban neighborhoods where nearby communities are more like to share the similarities. Therefore, identifying the spatial clusters is helpful for effective strategic planning.

Then, the exploratory analysis compares the characteristics of identified neighborhood typologies/clusters. This study compares the means and standard errors of the mean across the clusters. The results illustrate the social disparities and explore the reasons behind the facts, seeking actionable improvements for underdeveloped urban neighborhoods.

Table 1. Variables for Clustering Analysis

Variables	Description
Median Household Income (MHI)	The median household income of the zip code
Latitude	The latitude of the zip code centroid
Longitude	The longitude of the zip code centroid
Offices	Offices and/or other co-work places
Professional Services	Doctor's offices, legal services, etc.,
Retails	Supermarkets, department stores and/or other stores
Restaurants	Restaurants, cafeterias and/or other food-service outlets
Schools	Primary, middle, high schools and/or colleges and universities
Sport and Fitness	Basketball courts and/or community pool etc.
Leisure	Parks and/or recreational places
Civic Services	E.g., churches
Transportation	E.g., airports, railroads, etc.
Other Land Uses	Not in the previous nine categories

.

RESULTS

The results start with the descriptive statistics for each input variable and then illustrate the neighborhood clusters. Finally, the exploratory analysis compares the socioeconomics and land uses across the clusters.

Descriptive Analysis

Figure 1 plots histograms of input variables. All the plots generally show gamma or normal distributions, which indicates the quality of data being fairly good. However, it is important to note that some variables (i.e., *retails* and *restaurants*) may show a second peak after the main peak, implying multiple clusters may exist in the data.

Figure 1. Histograms of Input Variables
Note: Latitudes/ longitudes show the normal distributions, and they are excluded from the figure

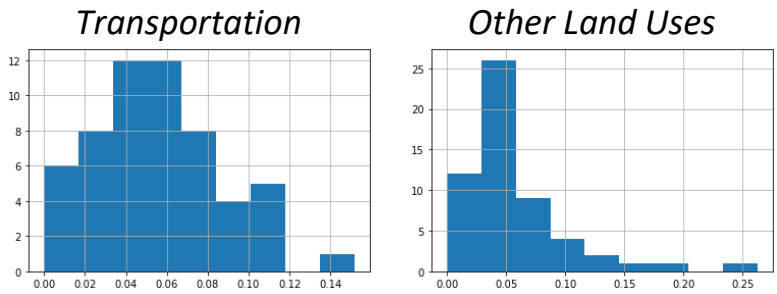
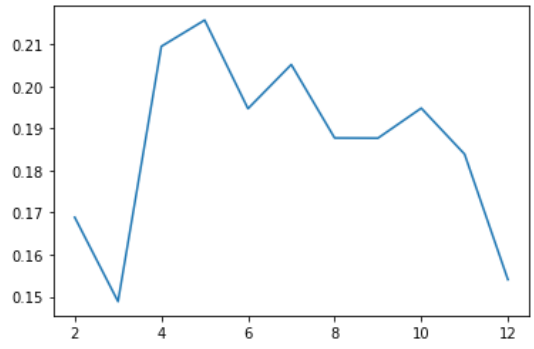


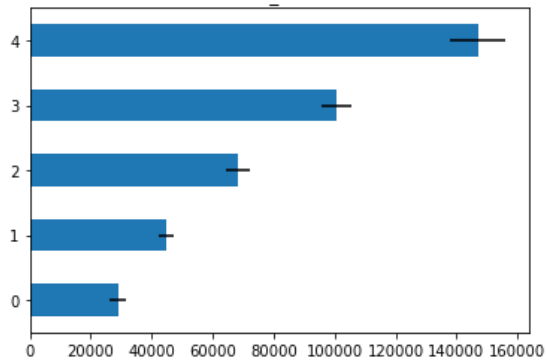
Figure 1. Histograms of Input Variables (cont)
Note: Latitudes/ longitudes show the normal distributions, and they are excluded from the figure

Neighborhood Clusters and Typologies

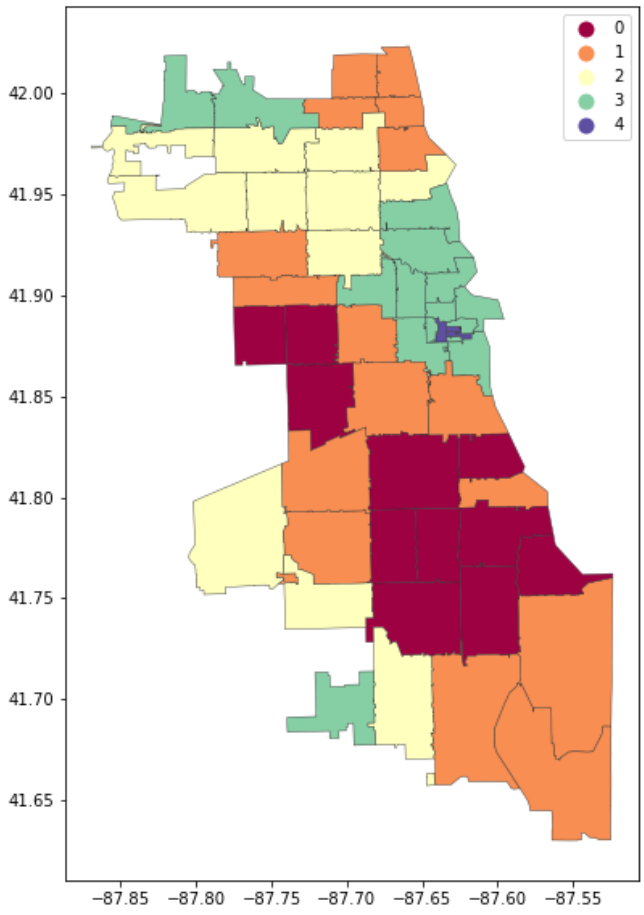
After standardizing the 13 input variables, the clustering analysis takes the standardized values and loops the k-means computation for *k = 2 to 12*. ***Figure 2(a)*** illustrates the Silhouette scores for these clustering results. The plot shows that the score reaches its first peak when *k = 5*. This indicates that all 56 Chicago neighborhoods are best to be clustered into five groups.



(a) Silhouette Scores for the Clustering Results (*k = 2 to 12*)



(b) Median Household Income of Neighborhood Clusters.
Note: error bars indicate 95% CIs (or 1.96 standard errors of the mean)*



(c) Neighborhood Clusters when *k = 5*

Figure 2. Results of Clustering Neighborhoods

Next, this study applies *k = 5* to categorize all Chicago zip codes and sorts the clusters by the average MHIs of classified neighborhoods. The error bars in ***Figure 2(b)*** suggest the average MHIs of these neighborhood clusters are statistically different. The colors of clusters, shown in the legend for ***Figure 2(c)***, present that dark red (or Group 0) is the neighborhood group with the lowest average income while dark blue (or Group 4) is the group with the highest average income.

Figure 2(c) also shows the distribution of neighborhood clusters and confirms with the grounded observations in the city. The highest income cluster (Group 4) locates in the city core, also known as the central business district (CBD). These neighborhoods are unique but very small in size. North to the CBD is historically affluent neighborhoods (Group 3). In contrast, the western and southern parts of the city are clusters of disadvantaged neighborhoods.

Exploratory Analysis

Clustering the neighborhoods is not intended to label the area with "good" or "bad"; rather, clustering analysis helps to visualize the social disparities, explore the underlying reasons, and probably identify strategies for improvements.

Figure 3 compares the characteristics of neighborhood clusters. As mentioned before, Group 4 is unique, especially regarding its land uses. On average, over 30% of venues within the neighborhoods are offices or co-worker places. Hence, other categories, such as Restaurants, Leisure (e.g., parks) are significantly lower in terms of the proposition.

As shown in ***Figure 2(c)***, most neighborhoods are fall in Group 0 to 3. As average MHI increases, the neighborhoods tend to have more office spaces, sport & fitness, and leisure places, while fewer retail and restaurants. However, for low-income neighborhoods (Group 0 & 1), it is evidence that they have fewer leisure places and sport & fitness venues. Access to outdoor spaces and activity places is essential for both physical and mental health. Inadequate investments in such infrastructure and lacks of related services further downplays the quality of life within those neighborhoods.

Moreover, the civic service venues are exceptionally high in Group 4 neighborhoods than in other neighborhoods. It is suggested the opportunities of utilizing existing infrastructure to provide services that are lacking in those areas.

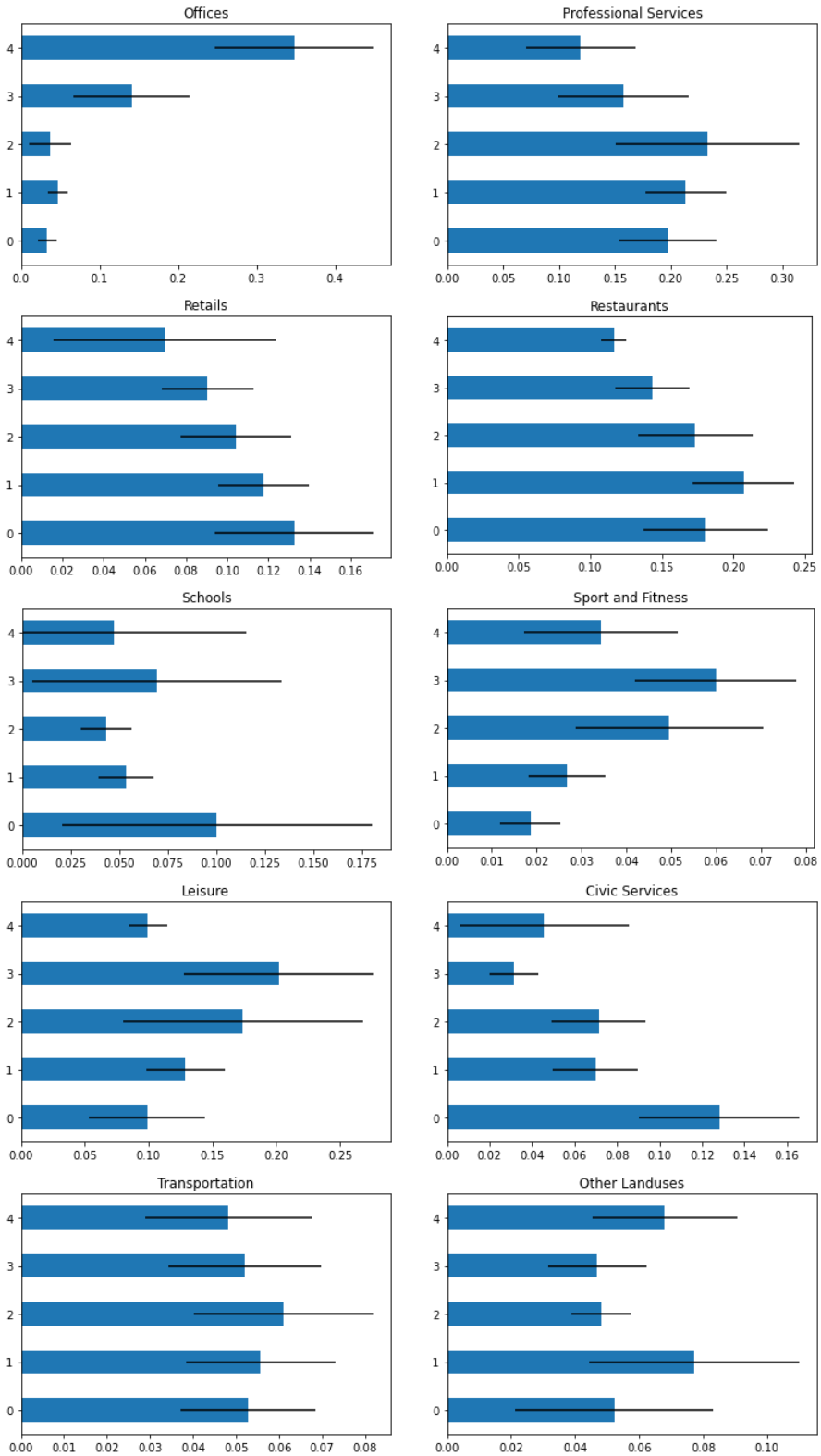


Figure 3. Comparing the Characteristics of Neighborhood Clusters
Note: error bars indicate 95% CIs (or 1.96*standard errors of the mean)

CONCLUSION AND DISCUSSION

This study collects data and identifies the typologies of Chicago neighborhoods based on socioeconomic factors and built environment features. All neighborhoods are then categorized into five groups. The results show that:

- the income disparities are also associated with varying numbers of infrastructure and services in the neighborhood,
- the sport and fitness venues, such as basketball courts and activity centers, are notably lacking in low-income neighborhoods,
- leisure places, such as parks, are also fewer in the historically disadvantaged area

The results suggest public and private investments pay more attention to these areas. Strategic planning and capital investments could narrow the gaps among neighborhoods and improve the quality of life for all urbanites.