Identifying Transit Deserts by Using Linear Regression and Clustering Algorithms

Emma Yumeng Wang

The Ethel Walker School, Simsbury, Connecticut

25ywang@my.ethelwalker.org

Abstract. As cities strive to increase sustainable transportation options, understanding and addressing transit deserts—areas where public transit is insufficient to meet residents' needs—becomes essential. This study examines transit deserts within Chicago by integrating sociodemographic data and public transit usage patterns. Through linear regression and clustering methods, key population characteristics influencing passengers' reliance on public transit across community areas are identified. Additionally, the analysis of Divvy bike usage data highlights disparities in bike station distribution, with most stations concentrated in central Chicago. This concentration limits transportation accessibility for outer areas, which may have latent demand for increased transit options. Our findings suggest potential high-demand areas lacking adequate service, supporting the case for a strategic redistribution of transit resources. The methodology and insights of this study extend beyond Chicago, offering a framework for identifying transit deserts in other urban centers to enhance equitable transit access and improve urban mobility infrastructure.

Keywords: Transit deserts, public transit usage, linear regression, clustering.

1. Introduction

Public transit is a cornerstone of urban infrastructure, influencing economic dynamics, environmental sustainability, and societal equity. Reliable access to public transportation is essential for the daily activities of millions, from commuting to work and school to accessing healthcare and other life necessities. People need public transit: Transit service provides people with a less expensive option to travel for daily purposes; sufficient transit service guarantees people's fair access to resources in the city. Public transit is essential for the equity of society, especially for people who do not have car ownership, which are defined as captive demand since they are closely related to job opportunities, groceries, and life essentials. A lack of transit service, on the other hand, limits people's access to not only basic amenities but also opportunities like jobs and education. In this case, resources will be exclusive to people with car ownership, which makes it detrimental to equity in general.

Areas where the supply of public transit doesn't meet the demand of the local population for public transit are called transit deserts, a term determined by Jiao in 2013 with a case study in four U.S. cities: Charlotte, North Carolina; Chicago, Illinois; Cincinnati, Ohio; and Portland, Oregon. Understanding and addressing transit deserts—areas underserved by public transit—is crucial for promoting urban mobility and sustainability. The definition of transit deserts draws attention to transit demand, an important piece in understanding the concept. Transit demand is closely related to sociodemographic traits: Areas with a high population generally have a higher need for transportation, and so some other cases like areas with high employment rate, which speaks for a high demand to commute to work instead of staying at home. Among all population groups, people with characteristics like low-income, young-age, old-age, disabled, etc., are more likely to rely on public transit, considering that they are less likely to own and use vehicles but still need to travel around in the city for daily purposes. Given this, transit deserts can be identified by recognizing places with higher demand for public transit, which are where people with those traits are more concentrated. Where there is high demand but the existing public transportation system doesn't keep up with it—for example, a sparse station coverage—can be then considered a transit desert. The consciousness to recognize transit deserts is necessary. Detecting and addressing the transit desert issue, as opposed to

indulging an insufficient public transit system to exist and putting the marginalized people more vulnerable, allows improvements in transportation in the areas that need fair access to transportation the most and helps achieve equity.

2. Literature Review

The concept of transit desert was introduced in 2013 by Jiao et al., who took the inspiration from the "food deserts," places where people don't have stable access of healthy food, and compared it to transportation, suggesting that places where their public transit supply doesn't meet people's needs to it are, similarly, transit deserts [1]. In the conversation of transit deserts, the transit accessibility and equity are accentuated. The importance of equity in urban planning and public transportation was especially addressed in their study: While the usage of vehicles can be different conversations among people—some are unable to afford vehicles and some are able to drive, it is necessary to make sure they are as flexible in traveling and have the same resources and opportunities as those who own vehicles and can drive. The equity in public transportation—that everyone has the same access to public transportation—ensures that not all transportation resources lean toward the advantaged groups while making the marginalized who are the most in need of transportation in an even more disadvantaged posture.

The study by Jiao et al. included previous research covers methods to measure transit accessibility and equity [1]. Jiao et al. proposed the transit desert as a concept that studies the gap between the transportation demand of a particular population and the public transit supply, they have access to [1]. It identifies that the transit-dependent population is usually marginalized people, which makes them vulnerable to the environment where the public transit service is too insufficient to meet their exceptionally high demand for public transit. Maharjan et al. recognized that transit accessibility is related to demographic factors, including vehicle ownership status, age, income, racial groups, etc [2]. Al Mamun further breaks down the measurement of transit accessibility into three components [3], which are trip coverage, spatial coverage, and temporal coverage. Currie identifies a new approach that uses socioeconomic statistics to measure the gap between transportation needs and supply [4]. These previous studies identified some important variables, such as average household income, average vehicle per household, which determine whether a population is marginalized and therefore reflect the need for public transit and their chance to access public transit.

However, most of these previous studies focused only on traditional public transit modes, such as bus and metro, but neglected the emerging public transit possibilities, such as shared bikes. To take a step further, our work focuses on demographic characteristics, to do the causal inference of the mode share usage in each census tract and the sociodemographic data. Clustering methods are used to identify people who depend on public transit services. The service of the emerging shared transit mode, like Divvy bike, is analyzed to examine the gap between transit demand and supply.

3. Methodology

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. Linear regression can be used to estimate the effect of an independent variable on a dependent variable, assuming that the relationship is linear and other assumptions are met. Linear regression analysis was used to determine the effect of different factors on one's choice to use public transportation. The equation for linear regression analysis was

$$Y = \beta X + C$$

Where Y represents the public transit usage ratio and X represents the vector of various features of the local population, such as household income, household size, and average vehicle number per household, etc. β is the coefficients that show the relation of the variable X and the target Y. Namely, with the increase of one unit of X, the result of target Y will change by β.

Decision tree is a machine learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure. The decision tree method sorted the priority of the features impacting the public transit ratio. It divided data into smaller sets according to specific features, which it determined were optimal split points. Eventually, all data were sorted under specific labels, and the most significant factors can be figured out.

Random forest is an ensemble learning technique for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or regression of the individual trees. The random forest method was a collection of decision trees, and the results were averaged together to give the quantitative importance of the features that could affect the public transit ratio. By these means, the coefficients and the significance can tell which factors are contributing to people's usage of public transit.

4. Data Exploration

The sociodemographic data comes from the American Community Survey (ACS) from 2017-2018. The sociodemographic variables used included total population, average household income, average vehicle per household, and employment status, and etc, as shown in Table 1.

Table 1. Statistics of Chicago Demographic Data

Variables	Description	Mean	Minimum	Maximum
households	Number of households in community areas	1349.32	113.0	12017.0
rent_median	Median rent	1134.42	274.0	2563.0
household_size_avg	Average household size of community areas	2.72	1.33	35.68
edu_bachelor_ratio	Bachelor ratio	0.22	0.0044	0.62
inc_median_ind	Median individual income	33970.00	4494.0	96667.0
employment_unemployed	Number of unemployment	926.14	97.0	9362.0

To further look into the emerging public-accessed transit mode, the Divvy Bike usage data from the Chicago Data Portal is also analyzed. The data contains a total of 36117 divvy sharing bike trips in the first week of January and a total of 154300 trips in the first week of August 2023, information including the locations of origin and destination, timestamps, and membership status. The origin and destination information are used to find the community area where and when the trips start or end. The membership status determines whether the bike user takes the route routinely or not. When categorizing the trips into 4 time slots—peak hours from 6 am to 9 am and from 4 pm to 7 pm, and non-peak hours from 9 am to 4 pm and 7 pm to 6 am—based on their time information, a lot fall into the peak hours.

5. Case Study

In the linear regression analysis, various features in each community area were used to predict the public transit ratio, which was then compared with the actual public transit ratio. The R^2 score was calculated to show how the predicted ratio is close to the ratio in the training set, and the R^2 score of approximately 0.86 indicated that they were quite close. The predicted and the actual ratios fit into a linear model, whose coefficient represented the importance of the features. Among features used, households, rent_median, household_size_avg, and edu_bachelor_ratio all gave a positive value, meaning that the increase in these features facilitates one's choice regarding the usage of public transit, as shown in Table 2. The features inc_median_ind and employment_unemployed, on the other hand, have a negative coefficient, meaning that their increase inhibits one's choice to use public transit.

Table 2. Regression results of travel mode share based on sociodemographic factors

	coef	Std err	t	P > |t|
const	-260.249	102.717	-2.534	0.011
pop_total	-0.1525	0.014	-10.751	0
sex_male_ratio	706.2608	184.495	3.828	0
age_median	-15.578	1.606	-9.704	0
households	-0.9679	0.026	-36.568	0
inc_median_ind	-0.0089	0.001	-8.49	0
rent_median	0.138	0.02	7.028	0
property_value_median	0.0003	9.53E-05	3.143	0.002
edu_bachelor_ratio	858.085	120.976	7.093	0
rent_median	0.138	0.02	7.028	0
household_size_avg	84.0639	10.71	7.849	0
employment_unemployed	-0.2118	0.036	-5.825	0
R-squared: 0.874				
Adj. R-squared: 0.873				
F-statistic: 554.2				

Clustering is also done to analyze various features' role in using public transportation. The decision tree is applied to sort out the priority of the features. The importance of each feature is found. Households has the most importance, and then edu_bachelor_ratio, age_median, pop_total, rent_median in descending order, as shown in Table 3.

Table 3. Feature importance of decision tree model

Feature	Importance
households	0.8105
edu_bachelor_ratio	0.1757
age_median	0.0080
pop_total	0.0030
rent_median	0.0022

The random forest is used for the same purpose. Important features found include households, edu_bachelor_ratio, pop_total, edu_master_ratio, in descending order, as shown in Table 4.

Table 4. Feature importance of random forest

Feature	Importance
households	0.7883
edu_bachelor_ratio	0.1306
pop_total	0.0441
edu_master_ratio	0.0130
household_size_avg	0.0086

Chicago's transportation in general puts weight on the central area along the lakeshore. The main public transportation services in the city include buses, rails, divvy bikes, and taxis. For methods with prescribed routes like bus and rail, given the long, narrow south-north geography of the city, the routes run through east-west, but the vertical routes don't always reach the very end regions, requiring people to take connection routes. This could make areas geographically marginalized, which sometimes could be sociodemographically marginalized at the same time, more vulnerable. For the transit modes without specified routes like divvy bikes, the situation is relatively more varied, but they still show a pattern that the stations are more concentrated in the central and the tourist spots further up. Due to the variations, this study chooses to focus particularly on divvy bikes.

When dealing with divvy bike trips data, we divide all trips in January and August 2023 into four time slots, with peak hours from 6 am to 9 am and from 4 pm to 7 pm, and non-peak hours from 9 am to 4 pm and also 7 pm to 6 am. In these two months we are focusing on, the demand for shared

bikes was usually higher during the peak hours than during the non-peak hours. We find that among the bike trips in the first week of January, 2023, during the morning peak hour from 6 am to 9 am, most started at the Near North Side and the Near West Side, where University of Illinois Chicago is located, and ended within these two community areas as well. During the evening peak hour from 4 pm to 7 pm, besides the Near North Side and the Near West side, another large number of the trips started at the Loop, which is right next to the Near West Side where University of Illinois Chicago is located and ended within the same area. Another relatively large portion of the trips started at the Near South Side, which is next to both the Loop and the Near West Side and ended at the Near West Side. A similar number of trips as those started at the Near South Side started at Logan Square and ended at Logan Square as well. During the non-peak hours, most trips still started at the Near North Side and the Near West Side, but those starting from Lake View, where some tourist spots are located, also stand out. For the end points, most of them still landed within the same areas. Most of the trips were relatively short, as they ended within the same community area as where they initiated.

The bike trips in the first week of August 2023 share a similar pattern of being short and not extending beyond the starting community area. During the morning peak hour from 6 am to 9 am, most happened within the Lakeview, Lincoln Park, Near North Side, West Town and Near West Side, and during the evening peak hour from 4 pm to 7 pm, most happened within the Near North Side, and then Lincoln Park, Near West Side, Loop, and Lake View. During the non-peak hours, the majority of the trips initiated in the Near North Side, Loop, Lincoln Park, Lake View, and Near West Side, and ended in the same or some very nearby areas.

The bike stations are the most concentrated in the Central that covers Near North, Loop, Near South, while a few spread out in the rest of the city. We visualize all the trips in two maps, one showing all the starting points and one showing all the endpoints. Both the starting points and the endpoints are concentrated in central Chicago, which covers the Near North Side, Loop, and Near South Side, and also around Lincoln Park, as shown in Figure 1 and 2. The relatively marginal regions—some among them are having a low average vehicle per household—are less covered by stations, which put the residents of these community areas who need public transit the most into the dilemma of not being able to easily travel around.

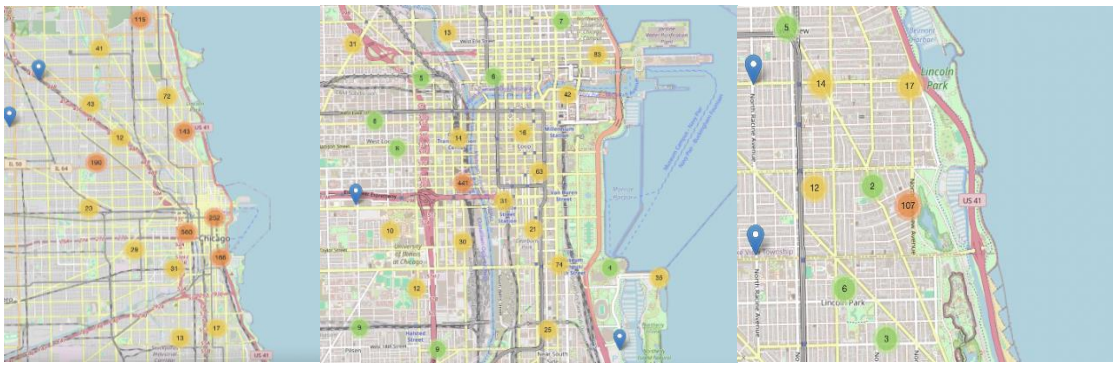

Figure 1. Ending sites

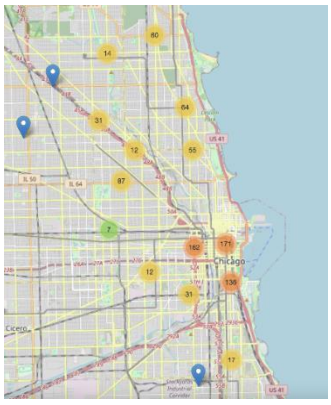

Figure 2. Starting sites

6. Conclusion

This study highlighted the critical issue of transit deserts within urban environments, with a focused examination on the city of Chicago. By integrating sociodemographic data and public transit usage patterns, our research has not only mapped out areas where transit services are insufficient but has also identified key population characteristics that predict reliance on public transit. The findings underscore the importance of considering diverse transit options, such as Divvy bikes, alongside traditional public transit modes to address urban mobility needs comprehensively.

Our regression analyses have shown that factors such as household size, income levels, and educational attainment significantly influence public transit usage. Particularly, areas with higher rates of vehicle non-ownership and lower income levels exhibited a greater dependency on public transit, aligning with the characteristics typical of transit deserts. The data-driven approach used in this study provides a robust framework for identifying underserved areas and suggests a redistribution of transit resources could significantly enhance transit equity.

Moreover, the utilization of emerging shared transit modes like Divvy bikes reveals gaps in current transit provisions and offers a lens through which to view potential enhancements in transit accessibility. Our findings suggest that expanding access to such shared modes in underserved areas could mitigate some of the challenges faced by residents in transit deserts.

In conclusion, addressing the issue of transit deserts is not only a matter of expanding transit services but also requires a strategic approach to ensure that these enhancements are equitably distributed. The methodologies developed and applied in this study serve as a foundation for urban planners and policymakers to implement more inclusive transit systems that cater to the needs of all urban residents, thereby fostering a more connected and sustainable urban future.

References

[1] Jiao, J., & Dillivan, M. (2013). Transit deserts: The gap between demand and supply. Journal of Public Transportation, 16 (3), 23-39.

[2] Maharjan, S., Tilahun, N., & Ermagun, A. (2022). Spatial equity of modal access gap to multiple destination types across Chicago. Journal of Transport Geography, 104, 103437.

[3] Al Mamun, M. S., & Lownes, N. E. (2011). A composite index of public transit accessibility. Journal of Public Transportation, 14 (2), 69-87.

[4] Currie, G. (2004). Gap analysis of public transport needs: measuring spatial distribution of public transport needs and identifying gaps in the quality of public transport provision. Transportation Research Record, 1895 (1), 137-146.