The Future of Chicago's Transit-Oriented Development: A Quantitative Analysis of Ridership by ACS and Zoning Variables

By Barrett Lopez

Submitted in partial fulfillment of the requirements for the degree of:
BACHELOR OF ARTS
IN GEOGRAPHICAL SCIENCES and ECONOMICS
at THE UNIVERSITY OF CHICAGO

Faculty Advisor: Dr. Crystal Bae
Environmental and Urban Studies Preceptor: Jamie Countryman

April 15, 2022

Table of Contents

Abstract

This thesis seeks to develop a comprehensive understanding of chronically underused stations throughout the Chicago Transit Authority's rail network. I will specifically focus on land use, demographic, and built environment indicators within block groups at a half-mile radius as independent variables, studying these variables with all rail stations in the CTA system from 2016-2019. Through the use of negative binomial multivariate regression models with American Community Survey (ACS) block-group level and City of Chicago Zoning Ordinance data, I seek to discern common trends shaping the current status of the system's use with a distinct focus on the large variety of ridership quantities in the system. Results show that demographic indicators of white population and employment, were significantly correlated with heightened ridership at a system-wide level. Significant zoning types offer unique contributions to ridership assumptions, as Downtown Core (DC), Private Development (PD), and Residential Single-Unit (RS) districts see estimated increased ridership proportional to their areas proximate to stations; the areas of Transit (T) and Neighborhood Commercial (C1) districts show negative estimates. Finally, the model suggests highly significant positive coefficients buildings finished within the eras of post-2000 and 1940-1969. Despite the existence of multiple significant regressors among each type of presumed transit effectors, high standard errors deter certainty of conclusions in each highly significant variable in the final model, which uses 35 unique regressors. This analysis should not be interpreted causally, but the recognition of significant zoning, demographic, and built environment coefficients in a suggest intricate relations to ridership at the system-wide level.

Introduction

Chicago's transit network is inextricably linked to the city's variety of neighborhoods, economic opportunities, entertainment options, and cultural vibrance. As a tool for granting access to this variety of opportunity, the CTA's rail options offer residents of all socioeconomic strata a cost-effective and relatively quick transit option. The ingrained nature of the CTA can be attributed to its substantive effect on the city's economic and social structures throughout the various reimaginings and expansions of the system, serving residents and tourists alike through distinct political, social, and economic eras of the city. Yet certain stations in Chicago show consistently low ridership metrics such that planners must consider their viability as they seek to facilitate this access in future planning with reallocations or influxes of funding. This consideration must account for the function of each station as a node within the CTA's vast transit network - a complex, idiosyncratic entity warranting further explanations than basic station to station comparisons.[1] To contextualize transit in Chicago requires acknowledgment of these differences within the system – a multiline station within the Loop, linked to nearly every other station in the city with relatively minimal effort, is nearly impossible to simply compare with a station in a highly industrial stretch of the Orange Line, a line intended to service the airport and introduce renewed technological and operational validity to the CTA.[2] Thus, comparison and interpretation warrants consideration of a variety of demographic, built environment, and zoning covariates that form the foundation of this study.

[1] Sabyasachee Mishra, Timothy F. Welch, Manoj K. Jha,
Performance indicators for public transit connectivity in multi-modal transportation networks,
Transportation Research Part A: Policy and Practice, Volume 46, Issue 7, Pages 1066-1085. 2012.
[2] Staff. "CTA Celebrates 25th Anniversary of Orange Line Service to SW Chicago & Midway Airport." *CTA*, CTA, 30 Oct. 2018, https://www.transitchicago.com/orange25/.

While its viability as a metric of accessibility is often debated, I will consider the standard half-mile[3] access radius from a station as the area of concern when collecting data and assessing independent variables. This is duly beneficial, as it allows for large-scale geographic analysis in case-study formats as well as generating useful statistics and analytical opportunities for the surrounding areas of each rail station. This is accomplished through the aggregation of block group data to each station on the condition of centroid overlaps, which will be further discussed along with other methodological considerations.

Despite extensive study of the CTA's operational failures, unsuccessful early transit-oriented development efforts,[4] and possible expansions,[5] contemporary transit literature is often strictly concerned with the CTA's system-wide successes or resulting land use. Thus, the common focus is rarely placed on the variety of spatial factors contributing to underuse of the transit stations, instead attempting to pinpoint the effects of transit systems on neighborhoods. While this approach is undoubtedly relevant for increased transit equity, I believe that acknowledging the result of continued spatial, social, and economic processes surrounding the stations in question can properly elucidate human elements of transit use. In reversing this approach to acknowledge the effects of neighborhoods on the transit system, I hope to add a nuanced approach to existing understandings and arguments surrounding transit-oriented development.

---

[3] Erick Guerra, Robert Cervero, and Daniel Tischler. 2012. "The Half-Mile Circle: Does It Best Represent Transit Station Catchments?" Transportation Research Record: Journal of the Transportation Research Board, 2276: 101–109.

[4] Center for Neighborhood Technology, 2013. "Transit-Oriented Development in the Chicago Region: Efficient and Resilient Communities for the 21st Century" https://www.cnt.org/publications/transit-oriented-development-in-the-chicago-region-efficient-and-resilient-communities

[5] Zotti, Ed, 2016. "The Case for Rail Transit Expansion in the Chicago Central Area" National University Rail Center – NURail.

This study does not directly seek to prescribe solutions to long-standing socioeconomic and transit inequities in the City of Chicago, but I do hope to address the variables associated with low ridership that might adversely affect the continuously underserved – and often how low transit use in these areas, as is the case of the Green Line's branches in both the West and South Sides[6] – can be seen in stark relief to more affluent areas when assessing the West and South sides of the city in comparison to the North Side (Figure 1).

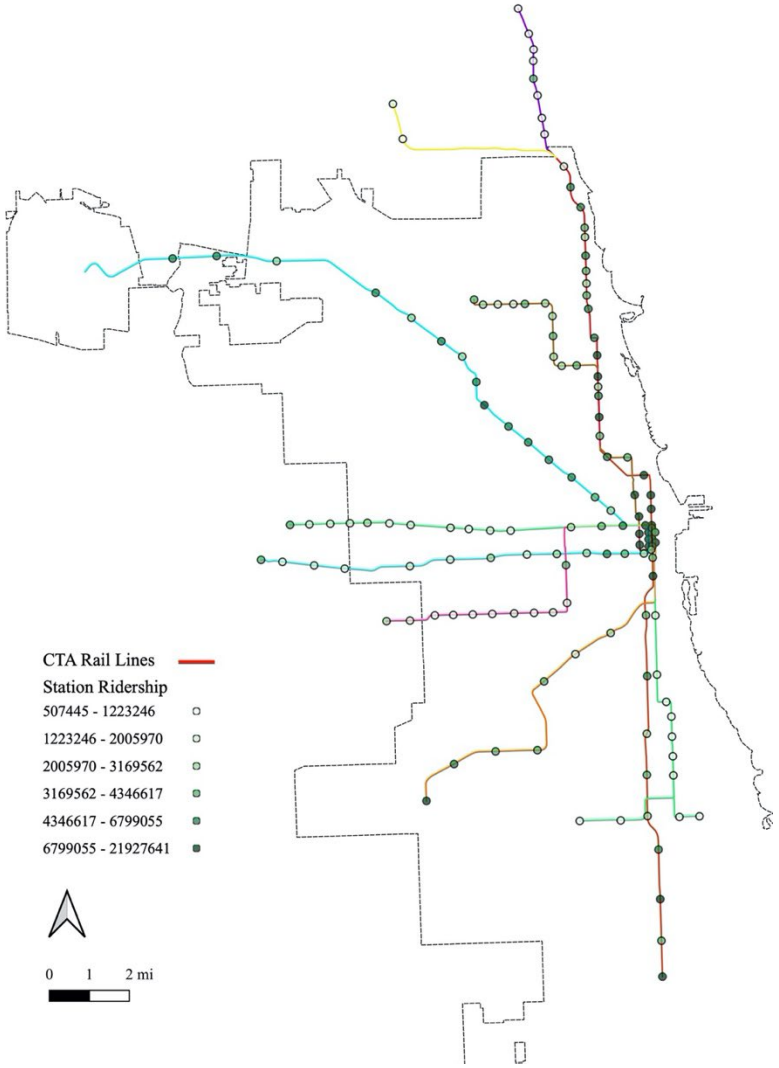

Figure 1: System Map of Ridership. Source: Chicago Open Data Portal

[6] Hertz, Daniel Kay. "Opinion: The Green Line's Waiting Game." *South Side Weekly*, 14 Nov. 2017, https://southsideweekly.com/opinion-waiting-game-green-line-cta/.

Literature Review

Introduction

A significant amount of past research, quantitative and qualitative alike, has paid due attention to the factors affecting ridership and transit access in Chicago and similar global cities. While much of this research uses dissimilar methodology and studies other cities, these relations and common understandings assist in shaping a comprehensive perspective ingrained in established qualitative and, most applicably in this case, quantitative theories and best practices surrounding transit study. The following subtopics and associated papers serve to ground this study in a historical understanding of transit theory and history in Chicago and at the national level. With regard to this study's intent to focus on modern issues while addressing well-established urban theories, a range of perspectives and issues are presented in the interest of alignment with best practice variable selections and assessment types, consideration of emerging factors, historical transit trends, and variable weightings due to unique historical factors affecting Chicago transit users' relation with the CTA rail system's effective area.

Distance Effect

Distance from public transit is among the clearest variables for use in understanding individual and thus community-wide incentives to use transit options. Lindsey et. al researched this phenomenon in the City of Chicago by considering both the Metra and CTA with direct focus on commuter transit use and its environmental effects, finding that commuters often highly preferred the use of private vehicles even when the transit stations were within a mile of their work destination.[7] The authors posit that car use in these cases is nearly unsolvable by transit

[7] Marshall Lindsey, Joseph L. Schofer, Pablo Durango-Cohen, Kimberly A. Gray,
Relationship between proximity to transit and ridership for journey-to-work trips in Chicago,
Transportation Research Part A: Policy and Practice, Volume 44, Issue 9, 2010, Pages 697-709. ISSN 0965-8564.

location alone and thus requires a degree of policy change, a recurring notion in transit literature.[8,9] Although the paper is generally untethered to the historical structuring of the city's various public and private transportation options, their findings elucidate the willingness of Chicagoans to use privately owned vehicles (POVs) as a preferred substitute to public transit in often inefficient situations.

Externalities of Public Transit Development

With the intent of provided comprehensive of analysis of the CTA's ridership issues, this paper must acknowledge the myriad use cases that exist in tandem with Chicago's variety of residents and economic opportunities. Thus, an understanding of transit's positive and negative externalities provide context for the choices of commuting residents at all scales. In "Evaluating public transit benefits and costs: Best Practices Guidebook," Litman elucidates these multifaceted relations with comparisons between public transit and highway infrastructure outcomes. Among his main evaluative claims is an overarching notion that transit infrastructure benefits should be logically supported even by non-users who reap the city-wide positive effects.[8] This is an extrapolation of his economic theory surrounding transit effects, in which "public transit and automobile transport have opposite cost curves" since "transit costs decline while automobile costs increase with density" (80). Thus, the implications of a growing city are directly aligned with the optimization of these costs: "transit service experiences scale economies" such that "transit improvements are often more cost effective than accommodating additional automobile travel on urban roads" (80). The aforementioned city-wide effects are notably applied to these automobile users and only becoming more relevant with his suggestion of decreasing cost efficiency of driving and thus the inverse occurrence in public transit

[8] Litman, T. (2021). Evaluating public transit benefits and costs: Best Practices Guidebook, Victoria Transport Policy Institute.

development. Litman cites factors including "aging population, rising fuel prices, increasing traffic and parking congestion, increasing urbanization, increasing costs to expand roads and parking facilities…and increasing health and environmental concerns" (80). This suite of benefits is inherently alluring to planners and all proponents of transit development in the face of typical urban highway infrastructure's dominating presence, yet in Litman's theoretical framework I note a lack of acknowledgment of the transit and automobile tradeoffs facing commuters in an ideal situation of well-developed and supported transit infrastructure. Despite mention of subsidized transit incentives (90), his suggestion of diminished city-wide health costs appears to be related to heightened incentives for drivers amidst lower congestion on roadways and ease of parking. Regardless, his notion of scaled benefits is relevant to the scale of this study given the neighborhood effects of train stations in the ideal, offering confounding variables of preferred location near transit due to associated benefits even without personal or familial use of the CTA.

Transit Oriented Development in Chicago: History and Perspectives

In Chicago, the existence and promotion of TOD policies have yielded less impact than optimistic politicians and planners hope for.[9] The Center for Neighborhood Technology, a Chicago-based group, assessed the Chicago metro area's development patterns in comparison with other large American cities such as Boston, New York, Philadelphia, and San Francisco, finding that "the rate of growth in the number of households was greater in the entire Chicago region than in Chicago's transit shed" which "contrasts with our peer regions where household growth occurred disproportionately around transit stations" (10). Boston and San Francisco, for instance, show high magnitudes of growth in their transit sheds at multiples of 2 and nearly 3.5

[9] "Transit Oriented Development in the Chicago Region." *Center for Neighborhood Technology*, 2013.

times that of their regions, respectively. The CNT partially ascribes Chicago's perceived developmental underperformance on "the Chicago Housing Authority's Plan for Transformation that eliminated 18,366 units in the City of Chicago," mainly old public housing projects under their purview, including 5,703 units that remained occupied and were thus counted in comparative measurements (10). Given housing unit growth of only 9,000 units in the study period of 2000-2010, this proportion is effectively diminished by the CHA's demolitions. Of further interest in the CNT's findings is the development patterns in the transit shed as of 2010, with a regional decrease in average household size "by about two percent while average household size in the transit shed decreased over five percent," presuming this change to be caused by the fact that "many TOD developments have featured small one- and two-bedroom condos" (12). This trend is reflected in a five percent decrease of family households in the shed paired with a six percent increase in non-family households in the same time period (13). To attribute ineffective TOD solely to provided housing types would be an oversight, but access for a wider range of families appears to be a lingering issue in defining transit users through typical urban housing options. While the CNT's findings convey the scale of these trends, their research does not provide the insight into the housing modes' effect on transit use other than brief mention of equitable access for families.

Evaluation of Transit Substitution Effects – Rideshares

Rideshare options present a confounding variable in transit use, operating as a unique substitute for privately owned vehicle use and public transit alike, especially in the commuter case. In "TNC use, Transit, and Vehicle Ownership in Chicago," the CNT uses quantitative evaluations to determine the association of rideshare use in Chicago with other variables, most

notably recognizing the effect of these options at filling mobility gaps.[10] The Center's findings include interesting conclusions surrounding the association between existing access and choice of rideshare utilization throughout Chicago from November 2018 to July 2019. The authors display a map of higher TNC use amongst a prevalence of transit throughout much of the higher income North Side, with a cumulative assessment that "the biggest predictor of TNC use is job and household intensity, but higher use is also correlated with demographic factors like household income and racial and ethnic compositions of communities." While this is relatively vague in the significance of these linear relations and lacks the similarity of scale to the composition of this thesis, it does warrant possible further weighting of income as a covariate in any regression models given higher willingness to pay for transit methods such as TNC that are unaccounted for by household vehicle ownership.

The relation between rideshare growth and transit use has been further investigated at a larger urban scale, with results from a range of American cities suggesting definitive correlations between rideshare introduction and modes transit ridership.[11] In "Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?" authors Graehler, Mucci, and Erhardt conduct an intensive quantitative study across twenty-two U.S. metropolitan areas, finding Uber's introduction to an urban area to often be a negative coefficient, but stating that "the commuter rail coefficient is positive, suggesting complementarity, but insignificant" (12). Other modes follow the overall negative trend with more statistical relevance, as "heavy rail and bus coefficients are negative and significant," which "suggests that TNCs reduce transit ridership" in a compounding fashion given that "heavy

[10] "TNC Use, Transit, and Vehicle Ownership in Chicago." *Center for Neighborhood Technology*. 2019. https://www.cnt.org/blog/tnc-use-transit-and-vehicle-ownership-in-chicago.
[11] Graehler, Michael & Mucci, Alex & Erhardt, Gregory. (2019). Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?.

rail ridership decreases by 1.29% per year, and bus ridership decreases by 1.70% percent per year" (12). Due to the lack of statistically significant relation between TNCs and commuter rail, the authors' assertion of complementarity requires further investigation; however, the suggestion could help to explain the muddled relationship as one of substitution in certain use cases where bus routes or alternative methods such as bikeshare are considered inefficient amidst a lower willingness to pay for full TNC trips, prompting riders to choose TNCs to increase access to commuter rail options. Whether this phenomenon sees the same growth as the compounding decrease in bus and heavy rail remains debatable, but the authors' quantitative findings suggest a uniquely positive compounding effect over time despite the caveat of statistical insignificance. As there is no existing data to directly account for block group level rideshare use – other than carpooling as a proxy – effects such as proliferation of ridesharing are considerable as omitted variables in this analysis.

Land Use Variable Effects: San Francisco Case Study

The nebulous idea of neighborhood structures invokes a range of land use types, prompting difficult constraints of variable inclusion. Prior work has attempted to quantify the effects of land use variables on transit use by utilizing comparative, matched case-study methods with areas of disparate urban structure.[12] In "Transit Choices in Pedestrian Versus Automobile Oriented Neighborhoods," authors Cervero and Radisch propose the validity of land use metrics via a transit-focused comparison of Bay Area neighborhoods proximate to Bay Area Rapid Transit (BART) rail stations. Rockridge, which the authors define as a "streetcar suburb," was among the first of the city's network of suburbs and attributed its quick growth to connection

[12] Cervero, R, Radisch, C. Travel Choices in Pedestrian Versus Automobile Oriented Neighborhoods. *UC Berkeley: University of California Transportation Center*. 1995. Retrieved from https://escholarship.org/uc/item/7cn9m1qz

with the trolley system. The town's land use is one of conscious mixed use, with a blend of retail and a dense combination of one- to four-unit housing options (14). The contrasting choice of Lafayette, a suburb characterized by post-World War II automobile-oriented planning with wide roads and a smaller retail core, is one of a town that experienced growth and establishment of identity in a different era of urban expansions less tethered to the Bay Area's rich transit history (14-15). The authors note a clear lack of mixed zoning and development, as zoning, namely retail, multi-family housing, offices, and single-family residences, sees clear divisions near the BART station and highway. The land-use disparity among these two suburbs appears to allow for worthwhile comparison in ridership given the relative controls of income and demographics.

In lieu of creating continuous variables associated with land use types by percentage, the authors instead create a dummy variable such that Rockridge = 1 while Lafayette = 0 (20). This methodology is logical in the case of the suburbs given their distinguished characteristics earlier in the paper, but the question of scale supersedes that of immediate application to other cities and transit systems. This also fails to address the specific land use types with positive correlation to higher ridership, only suggesting their relation by proxy with the high coefficient (.8291) attributed to Rockridge by way of multivariate logit regression (20).

Scaling Cervero and Radisch's work to Chicago could include the use of dummy variables to identify transit areas with similar trends, but I posit that the high possibility of colinear relations between predefined subtype binaries along land use lines would serve to muddle otherwise valuable conclusions and prevent the identification of useful land use indicators. However, their methodology holds value in its method of comparison between similar areas, if only at smaller scales.

Land-Use Variable Effects: Seoul

In "Effects of Land-Use Characteristics on Transport Mode Choices by Purpose of Travel in Seoul, South Korea, Based on Spatial Regression Analysis," Min, Lee, and Kim create a more robust model than the San Francisco study by incorporating spatial lag and autocorrelation techniques with the intent of determining the need for spatial regressions in evaluations of land-use effects on travel modes.[13] This study offers a city-wide perspective on land-use correlation analysis, and the use of spatial autocorrelation and spatial regressions provides an intensive method of assessing land-use effect across nearly all modes of transit. While the scope of my work, especially given use of relatively small half-mile buffers and incomplete data surrounding choice of commute type limits the ability to waver from use of basic OLS regressions towards spatial autocorrelation and regression, the validity of their techniques is worth consideration as a polar case to that of Cervero and Radisch. Their use of gross floor area (GFA) by land-use type (in this case only residential and commercial density) accounts for variations within each type while avoiding the use of unnecessary complex covariates in the model (4). Combining both measures into a residential-non-residential ratio to best suit their multivariate spatial analyses resulted in a finding that "planning with a similar scale of residential and non-residential buildings could be a factor that contributes to decreases in the use of public transit" (16). The use of a ratio appears highly applicable in Seoul, a city with a highly dense built environment compared to that of Chicago's, but could require tweaking if applied to Chicago. This also fails to account for attitudes towards transit given cultural and economic differences, among other more qualitative elements. Regardless of these location-based caveats, Min, Lee, and Kim's findings pose an interesting relation between the two land-use types such that the relation

[13] Min, B., Lee, G., Kim, S. Effects of land-use characteristics on transport mode choices by purpose of travel in Seoul, South Korea, based on spatial regression analysis. Sustainability (Switzerland), 13 (4), art. no. 1767, pp. 1-22. 2021.

between residential-commercial land-use ratios and public transit are deserving of further explanation. While I fail to have access to square footage metrics, which I believe would greatly increase the effectivity of multi-unit zoning types, their conclusions find validity as a conceptual foundation for the multivariate model I implement.

Demographic Factors and Considerations: Age, Race, and Others

Consideration of age provides context for common ridership groups within the CTA's coverage. The independent variable of ages 18 to 44 serves as a proxy for the working population most likely to necessitate trips to work; however, further understanding of age group preferences is required to discern attitudes towards transit by both generational attitudes and economic distinctions. Brown et al. address the state of ridership preferences by age group in "A Taste for Transit? Analyzing Public Transit Use Trends among Youth," discerning a higher likelihood for young people to use public transit, which they attribute to three integral factors related to age after studying transit use from 2001-2009.[14] In no order of importance, these include "life cycle factors common among young people (such as being a student, not yet having children, having a lower income)," "demographic factors (such as being a racial or ethnic minority)," and locational factors (such as living in densely developed, transit-rich neighborhoods)" (62). Notably, the authors do not attribute the variables of age or generational identity alone to any statistically significant correlations to transit use (61), offering validity to their identification of the three aforementioned factors with an appropriate selection of control variables.

Although Brown et al. direct their research focus towards age effects of transit use, the study's inclusion of racial and ethnic transit preferences provides nationally relevant statistical

[14] Brown, Anne E., et al. 2016. A Taste for Transit? Analyzing Public Transit Use Trends among Youth. Journal of Public Transportation, 19 (1): 49-67.

results. Using White, non-Hispanic riders as a baseline, the study found positive and highly statistically significant coefficients for ridership among minority groups. Given their use of a logistic regression to compare with the baseline white non-Hispanic group, the results convey a positive .96% difference in transit use among Black residents, a positive .66% among Hispanic residents, and a positive .65% difference for all other non-White non-Hispanic residents (59). Their incorporation of this finding into applicable conclusions may have repercussions for Chicago and the CTA, as they claim that continued transit use with age by the younger range of residents in these minority groups could lead to upward trends in future results associated with age (62). While the time period of data used in this study is over a decade old at its most recent range of 2009, the identified trends still provide valuable insight into subgroups that will serve as regression variables of significant interest in this study, albeit with the disparate intent of correlating system, line, and station ridership to these variables instead of assessing group tendencies of public transit use. Other useful metrics provided in the study include higher ridership among students and employed residents (59). While these are less permanent determinants, especially in the case of employment as a constantly shifting metric given macroeconomic trends (60), their inclusion in the Brown et al. study did lead to determination of significant positive coefficients in the OLS model at .51 and .32, respectively.

Zoning Ridership Effects: L.A. Study

In "Does Zoning Help or Hinder Transit-Oriented (Re)Development?" Schuetz et. al analyze the changes occurring within newly transit accessible Los Angeles neighborhoods.[15] As the basis of the study, Los Angeles' lack of well-established rail transit allows it to be a prime candidate for testing the causal effects of zoning. Of course, Chicago's transit history is

[15] Schuetz, Jenny, Genevieve Giuliano, and Eun Jin Shin. "Does Zoning Help or Hinder Transit-Oriented (Re)Development?" *Urban Studies* 55, no. 8 (June 2018): 1672–89. https://doi.org/10.1177/0042098017700575.

misaligned with the direct application of Los Angeles' above results due to the lengthy history of industrial and passenger rail systems in the area predating the CTA.[16] This fact, coupled with over one and a half decades between this study's focus and my own, bring into question the relevance of application to contemporary Chicago. Despite the incongruities between the two cities' built environments, I find validity in the conceptual framework of the authors' assessments of Los Angeles' current and future transit use; Schuetz et. al assess the case of L.A.'s public transit use with compatibility to transit-oriented development in mind by focusing on the more granular existence of nearby developments.

The authors make interesting conclusions as to the variety of development that emerges from zoning: "The form and timing of redevelopment reflects land values and zoning, as well as other public-sector actions – or lack of actions – around stations." As for these actions, which are inherently difficult to account for in a quantitative method, Schuetz et. al claim that "political support from neighbourhood residents and/or within-city governing bodies matters" when rezoning to allow for more density near stations. Addressing Chicago at the case study level is not the aim of this study, but Schuetz et. al make valuable conclusions as to Los Angeles' complexities in allowing for helpful zoning in a manner applicable to any urban space.

Negative Binomial Regression for Count Variables

In *Negative Binomial Regression,* Hilbe details the validity of Negative Binomial Regressions for use in research contexts including count variables, specifically in use cases where the data is Poisson overdispersed.[17] The quality of the model is derived from accounting for overdispersion, or a higher variance than mean among the dependent variable, which

[16] National Museum of American History Behring Center, Smithsonian Institute. "Chicago, the Transit Metropolis," https://americanhistory.si.edu/america-on-the-move/essays/chicago-transit-metropolis.
[17] Hilbe, Joseph M. *Negative Binomial Regression*. Cambridge: Cambridge University Press, 2007. doi:10.1017/CBO9780511811852.

invalidates the use of a Poisson model, the foundation of the negative binomial's framework. This is seen in Hilbe's description of the main adaptation of the model:

> *"The original manner of expressing the negative binomial variance clearly shows this mixture relationship: μ + μ2/v, where μ is the Poisson variance and μ2/v the two-parameter gamma distribution variance. We inverted the gamma scale parameter, v, to α, producing the negative binomial variance, μ + αμ. This form of the variance provides a direct relationship of α to the amount of overdispersion in the otherwise Poisson model" (219).*

The intricacy of this model arrives with the inclusion of the fitting or "overdispersion parameter," which properly accounts for the overdispersed data, allowing for more accurate standard errors given proper parameters for dealing with the otherwise unexpected dispersion detailed above. Later in the paper I refer to α as θ to match the model output in R, but the parameters are nonetheless identical. Hilbe's negative binomial model, the implementation of which is further discussed in the 'Regression Design' section, greatly increases the statistical viability of this study through log standardization and the introduction of the overdispersion parameter.

Conclusion

A breadth of transit research has been conducted to discern economic attributes of public transit as well as trends in use based upon variables such as land-use, socioeconomic factors, demographic factors, and vehicle ownership. Despite a breadth of technical variety amongst the papers cited in this review, the methods presented served to inform my own. As using proper techniques is integral in nearly all facets of this study – from data organization, manipulation, and selection to regression design and spatial considerations – the papers above set valuable precedent for positive and negative cases of analyses.

Even with the admirable scope of transit research completed in the last three decades due to ever-increasing availability of data, rail transit appears deserving of more focus. The interlinking of bus and rail options warrants assessments of whole systems, but I note the difficulty of discerning commuter rail use in multiple of these studies. Along with this, larger analytical concerns exist upon consideration of differing use cases between daily bus and rail use, especially for commuters. The subjectivity of studying single cities displays the nuance that must be applied to Chicago with historical groundings in case studies and possible variable weightings dependent on exploratory data analyses. This study hopes to address missing links in the current literature with a near singular focus on the CTA's commuter rail options while assessing direct influences upon ridership in multiple areas of the city at large, block-group level geographic scales uncommon in current literature.

Methodology/Analytical Plan

Demographic Data

In an effort to utilize relevant and recent demographic data, I utilize 2015-2019 Block Group data retrieved from the IPUMS National Historical Geographic Information System (NHGIS)[18]. I utilized the following 2015-2019 American Community datasets summarized at the block group level:

NHGIS Block Group Level Demographic Datasets:

B01003: *"Total Population"*

B25001: *"Housing Units"*

B25004: *"Vacancy Status"*

B25034: *"Year Structure Built"*

B08141: *"Means of Transportation to Work by Vehicles Available"*

B02001: *"Race"*

B19013: *"Median Household Income in the Past 12 Months (In 2019 Inflation-Adjusted Dollars)"*

B23025: *"Employment Status for the Population 16 Years and Over"*

[18] Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 16.0. Minneapolis, MN: IPUMS. 2021. http://doi.org/10.18128/D050.V16.0

Spatial Data

The following GIS datasets were utilized to create geographical models of the CTA Rail system, and later for joining and overlap analysis to generate independent regression variables:

Data	**Source**
CTA – 'L' Rail Stations	Chicago Open Data Portal
2016-2019 ACS Block Groups	NHGIS
Boundaries – Zoning Districts (current)	Chicago Open Data Portal

Ridership Data

Daily ridership data from January 2001 to November 2021 organized by station and line was acquired from the Chicago Open Data Portal.[19] As discussed in the following sections, ridership data is used as a dependent variable to understand overall ridership trends and determine common metrics for all stations in the system. To generally match the time period of the data used in the regressions, the ridership was summed across the 2016-2019 period of the ACS while still occurring after the implementation of the City of Chicago's Transit-Oriented Development policy.[20] This has the dual benefit of analyzing ridership prior the COVID-19 pandemic, which saw drastic decreases in CTA ridership amid system closures during early-to-mid 2020, and showing ridership mainly concentrated after the implementation of the transit-oriented development policy in late 2015.

[19] City of Chicago. CTA - Ridership - 'L' Station Entries - Daily Totals: Chicago Open Data Portal. https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f.
[20] Metropolitan Planning Council. "Chicago's 2015 TOD Ordinance," https://www.metroplanning.org/work/project/30/subpage/4.

Zoning Data

All zoning data was sourced from the Chicago Open Data Portal, using the most recently published boundaries from 2016.[21] Due to the extensive categorization of zoning due to ordinances and the prevalence of large, uniquely zoned private tracts throughout the city, zoning classes were partially consolidated for ease of consideration in the analysis. Certain zoning classes in Chicago have distinct types – especially those with mixed-use commercial, retail, and residential buildings, which fall under various classifications. Business district (B) zonings range from retail options on small neighborhood roads (B1) to full retail districts on high-traffic streets (B3). Subtypes, mainly for legislative detail, exist among each of these classifications. For example, B3 includes five subtypes (B3-1, B3-1.5, B3-2, B3-3, and B3-5) delineated by floor area ratios and lot area per unit. These subtyping standards apply to all other zoning categories seen below to varying degrees. While the presence of high or low square-footage constraints for businesses is of minimal concern in this analysis, a valid argument for more granular zoning covariates would be that other zoning classes, such as residential, have zoning types that would indicate socioeconomic conditions. I contend that other demographic covariates partially account for this relationship such that the presence of large houses in the generally dense built environment of the city would already be associated with a high presence of residential zoning and higher median income per household, another included regressor in the full model.

Downtown zonings, which include Downtown Residential District (DR), Downtown Mixed-Use District (DX), Downtown Core District (DC), and Downtown Service District (DS). While these are relatively similar to other zonings in their built environment and mostly determined given their centrality, I chose not to convert them to other zonings – such as

[21] City of Chicago. Boundaries – Zoning Districts. Chicago, IL: Chicago Open Data Portal. https://data.cityofchicago.org/Community-Economic-Development/Boundaries-Zoning-Districts-current-/7cve-jgbp

Residential Multi-Unit (RM) for Downtown Residential District (DR) – due to the extremely high ridership seen in the Loop area where these zonings are concentrated. As such, these zonings are utilized to duly convey and separate the effect of the downtown area's multi-line stations and the variety of non-commuter use cases for this area. This differentiation will be further analyzed in the results, but it first serves as a pragmatic control to increase goodness of fit given the high ridership seen in all of the Loop area stations.

Zoning Regressors[22]

Code	District	Description
RS	Residential Single-Unit District	Detached, single family homes.
RT	Residential Two-Flat, Townhouse and Multi-Unit District	Two-flats, townhouses, low-density apartment buildings, single family homes.
RM	Residential Multi-Unit District	Medium to high-density apartment buildings. Two-flats, townhouses, and single-family homes are also allowed.
B1	Neighborhood Shopping District	Retail storefronts on low-traffic streets. Apartments allowed above the ground floor.
B2	Neighborhood Mixed-Use District	Retail storefronts, apartments allowed on the ground floor. Intended to spur development in commercial corridors with low demand for retail.
B3	Community Shopping District	Shopping centers, large stores, and retail storefronts, often along major streets. Allows more types of businesses than B1 and B2 districts.
C1	Neighborhood Commercial District	Retail storefronts. Allows more business types than B1 districts, including liquor stores, warehouses, and auto shops.
C2	Motor Vehicle-Related Commercial District	Shopping centers. Allows more business types than B1 districts, including liquor stores, warehouses, and auto shops. Apartment allowed above the ground floor.
C3	Commercial, Manufacturing, and Employment District	Businesses and factories, no housing allowed. Serves as a buffer between manufacturing and residential/commercial districts.
DR	Downtown Residential District	High-rise apartment buildings, largely in the Gold Coast. Ground-floor stores are okay, offices aren't.

[22] All zone types and descriptions directly sourced from Second City Zoning

DX	Downtown Mixed-Use District	Downtown high-rises - offices or apartments - with ground-floor stores. Prevalent on the edges of Loop: east of Dearborn Ave, in River North, the South Loop, and the West Loop.
DC	Downtown Core District	High-rise Loop office buildings. Also covers downtown stores, entertainment, and civic buildings. Allows residential buildings.
DS	Downtown Service District	Rail yards, warehouses, and small businesses on downtown's periphery.
M1	Limited Manufacturing/Business Park District	Light manufacturing, warehouses, and wholesalers.
M2	Light Industry District	Moderate manufacturing, warehouses. Also allows freight and recycling facilities.
M3	Heavy Industry District	Heavy manufacturing, warehouses, and waste disposal - junkyards, landfills, and incinerators.
PD	Planned Development	Tall buildings, campuses, and other large developments that must be negotiated with city planners. Developers gain freedom in building design, but must work with city to ensure project serves and integrates with surrounding neighborhood.
PMD	Planned Manufacturing Districts	All kinds of manufacturing, warehouses, and waste disposal. Special service district - not technically a manufacturing district - intended to protect the city's industrial base.
POS	Parks and Open Space	Chicago's major parks, including Lincoln Park, Humboldt Park, and Washington Park.
T	Transportation	Bits of land designed to protect roads, bus ways, bike trails, and rail lines.

Figure 2: Zoning Regressors

*Non-Zoning Regressors**

Variable	**Type**	**Source**	**Category**
Ridership	Dependent	Chicago Open Data Portal	Continuous
Total population	Independent	NHGIS (ACS)	Continuous
Vehicles per capita	Independent	NHGIS (ACS)	Continuous
Median household income	Independent	NHGIS (ACS)	Continuous
Educational Attainment (% of pop. with bachelor's degree or greater)	Independent	NHGIS (ACS)	Continuous
White	Independent	NHGIS (ACS)	Continuous
Black	Independent	NHGIS (ACS)	
Asian	Independent	NHGIS (ACS)	
Housing Units	Independent	NHGIS (ACS)	Continuous
Vacant Housing			
Structures Built pre-1940	Independent	NHGIS (ACS)	Continuous
Structures Built 1940-1969	Independent	NHGIS (ACS)	Continuous
Structures Built 1970-1999	Independent	NHGIS (ACS)	Continuous
Structures Built 2000-2016	Independent	NHGIS (ACS)	Continuous

Figure 3: Demographic and Built Environment Regressors

Spatially Constraining and Aggregating Data in QGIS

To determine which block groups fell within the half-mile buffer range of the study and subset associated data appropriately, I utilized QGIS. I first created centroids for each block group in the city. To begin the process, I joined all ACS data to the block group geographies in the software using the unique block group identifiers to avoid any mismatches or overlapping data with less accurate identifiers. While I recognize the shortcomings of using centroids to represent a whole block group – especially in block groups with residential clusters representing the majority of demographic variables in the area – this method was chosen due to the unavailability of more precise block-level data and in lieu of over-counting small block group segments. I created half-mile buffers around each station, making the choice not to dissolve any buffer intersects between multiple stations. After the creation of centroids and buffers, I added a dummy variable column to the data with a default value of "0." With this dummy as a binary to indicate if a block group centroid falls within station buffers, I used the built-in "select point-in-polygon" function and assigned all selected points a "1" value.

To account for the spatial intricacies of zoning and the lack of geographic matches with block group designations, I made multiple considerations. Despite the possibility of heterogeneous use cases within zoning areas and types, each zone was considered uniform by class such that area per zoning per block group was assigned an area value in the data, by default square meters. Earlier iterations of the study were planned around equal consideration of each block group such that ridership per station would be considered equal in each block group. While the assumption of ridership as geographically constrained to a half-mile is inherently a proxy to more extensive networks of transit use, the equal distribution of ridership biased coefficients upward in demographically and spatially disparate block groups. Given this issue, I instead chose

to aggregate all block group values to one summary row/observation per variable per station station, retaining the same variables through use of weighted means and summations in R.

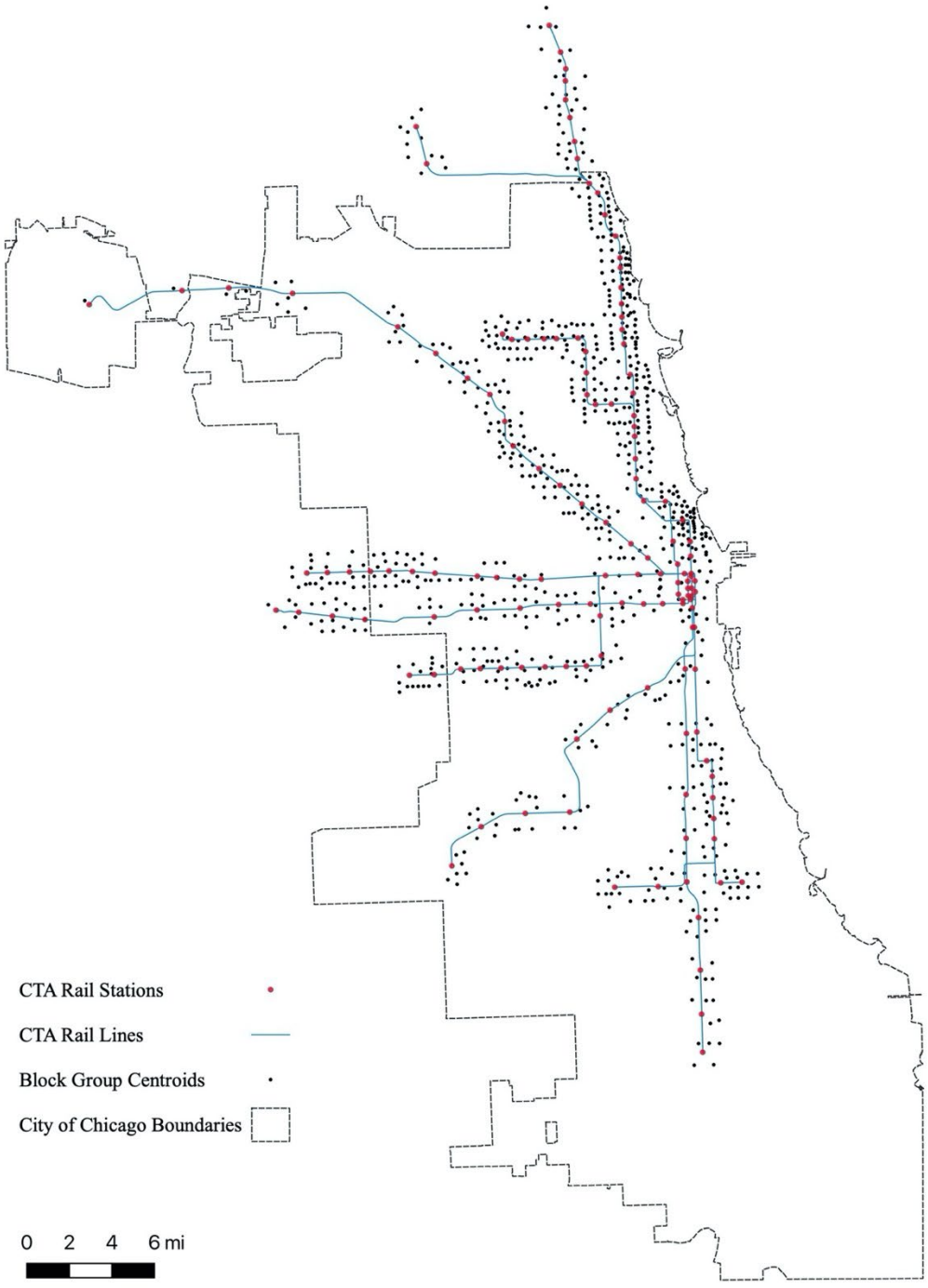

Figure 4: CTA Rail Lines, Stations, and Proximate (.5 mi) Block Groups. Source: Chicago Open Data Portal

A Brief Overview of Ridership, 2016-2019

At a system-wide scale, the CTA sees wide station-to-station variations in ridership throughout this period:

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
507,445	1,608,795	3,173,428	4,269,536	5,587,264	21,927,641

Kostner, a Pink Line station near the western extremity of the line, is the lowest in total rides in this time period with 507,445. By the same measure, the most 'utilized' station is State/Lake on the Red line with 21,927,641. This variation is relevant, but State/Lake's placement within a commuter-centric area of the Loop renders it useful for a range of weekday users of various economic backgrounds – a stark contrast from the manufacturing adjacent, single-family residential area surrounding the Kostner station.

In the interest of designing a model to account for variations in station locations and the ridership they commonly service, I utilize data to cover a wide range of demographic, land use, and built environment variations proximate to the system.

Data Visualization and Exploratory Analysis in R

After completing necessary cleaning and organization of the data, I used RStudio to determine overarching trends in variables of interest in an effort to contextualize city-wide metrics and prepare data for multivariate regression analyses. As I set out to understand the effect of a range of variables possibly affecting the system's function in both indirect and direct means of access, incentives, and transit substitutes, these early steps provided a means of checking data viability and comprehending possible limitations to the study. This exploratory data analysis included basic statistical visualizations and determination of missing data. Changes made to the data at this stage included the selection of all weekday dates as to best address

commuter rides and avoid outlying weekend and holiday ridership values in the model. The most general trend in question is that of ridership throughout the period, in which a decreasing trend is observed accompanying clear seasonality repeated each year:

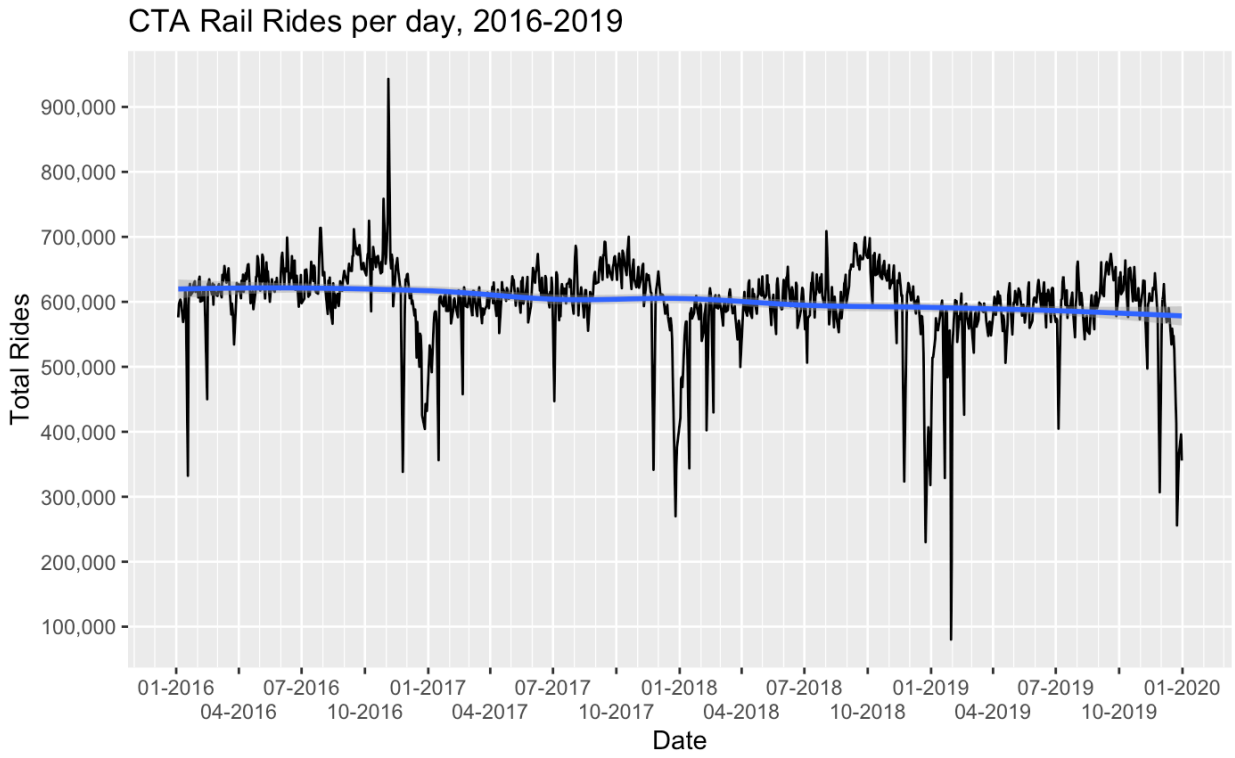

Figure 5: Total CTA Rail Rides by Day, 2016-2019

R was also utilized for time-series analysis, which assisted in recognizing 'random' (i.e. non-seasonally associated) ridership trends in the time-period of the study. This assisted in determining stations of interest for further statistical analysis when selecting case study stations in differing city regions, as any non-cyclical fluctuations or consistently low ridership compared to other line stops warrants further quantitative and qualitative analysis to determine exogenous neighborhood or city-wide effects. The time series analysis confirmed seasonality assumptions suggested by Figure 4, displaying relative consistency of seasonal trends and thus justifying the summation of ridership values per station as a basis for analysis and regressions. As seen in Figure 4, significant decreases in ridership are associated with the late-December to early-

January holiday season as well as inclement weather, such as the polar vortex which occurred at the era's ridership low-point on January 31st. This general uniformity and an observed lack of randomness in the time series, combined with the lack of time series data for demographic and zoning further warranted summation.

As prerequisite to regression design, I chose to investigate the distribution of ridership values. Figure 5 displays the non-transformed distribution of ridership among stations, while Figure 6 displays the log-transformed distribution:

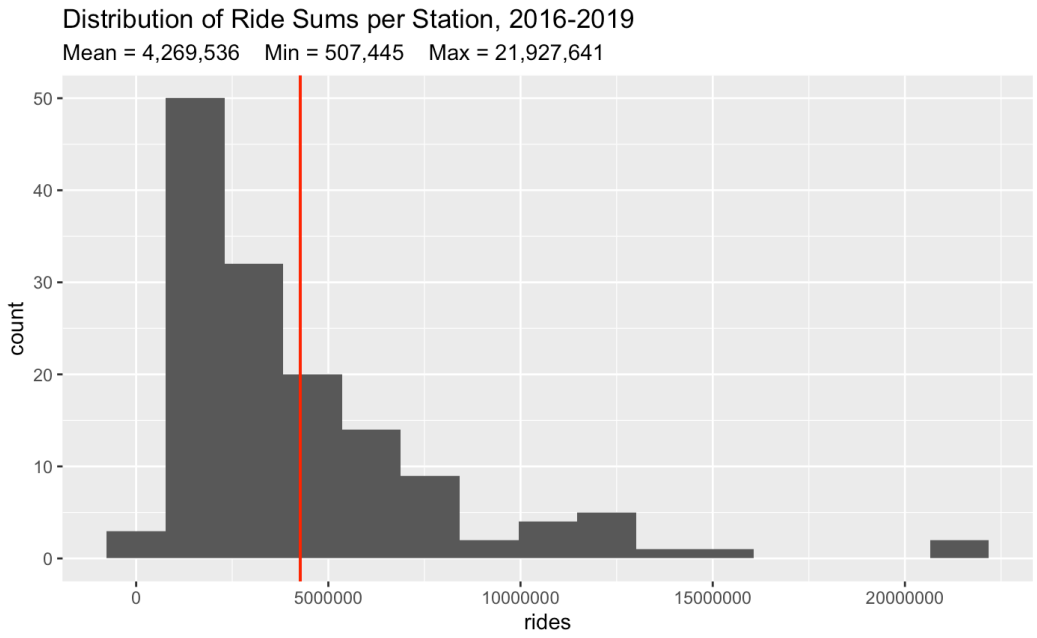

Figure 6

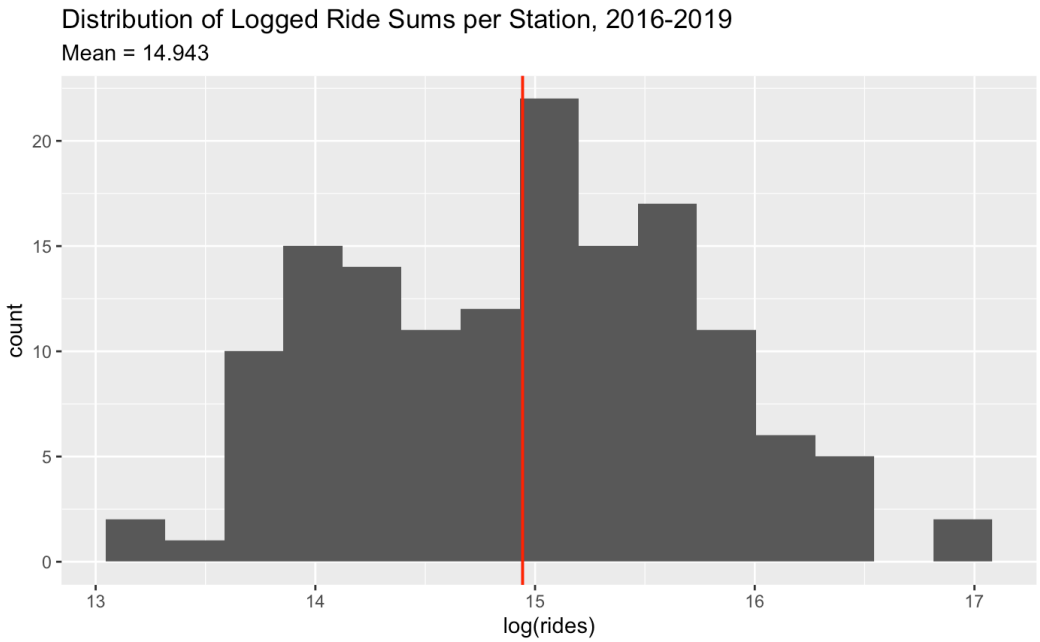

Figure 7

As visible in the above figures, the non-transformed distribution (Figure 5) is right skewed with notable high ridership outliers, while the log-transformed distribution (Figure 6) displays general lognormality despite the presence of partial upper-quantile skew due to high-ridership stations.

Importantly, I also analyzed the data with regard to modeling choices. This mainly takes place in the form of determining the existence of overdispersion, or larger variance than mean amongst the dependent variable in the observations. As shown above, mean ridership is equal to 4,269,536; observed variance is 13,923,261,082,479 and thus massively over-dispersed relative to the mean. This disqualifies the use of Poisson models and requits more intensive logarithmic modeling options to fit accordingly in the presence of overdispersion.

Summation and Weighting in R

The majority of variables were grouped and summed by station, which created station-based rows for all proximate block groups. Percentage variables from pre-existing block group data were weighted by their universes; for example, race percentage variables were weighted by

total population per block group to create an accurate average per station area. The majority of these variables were only utilized in the exploratory process, as raw counts were preferred in the construction of the final model due to modeling choices and ease of eventual interpretation.

Regression Design

Addressing Over-Dispersion in Modeling: The Negative Binomial Model

Considering the use of a count variable, I hoped to use a Poisson regression to assess effects; however, the observed overdispersion led to the use of a Negative Binomial model instead of Poisson or Quasi-Poisson options, the latter of which is also commonly used to evaluate over-dispersed data.[23,24] This choice arrives with recognition of the block group data's spatial groundings, as the negative binomial model provides more equal weighting to smaller sites and allows them more significant effect on the model (Ver Hoef and Boveng, 2007), granting stations with fewer block group centroid overlaps relative equality to those in densely populated and thus centroid-dense regions. The implementation of a negative binomial model serves to best replicate the Poisson model in this situation of large-scale overdispersion. I choose to utilize a system of three negative binomial models, the results of which are visible in Table 1.

[23] Ver Hoef, Jay & Boveng, Peter. (2007). *Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data?.* Ecology. 88. 2766-72. 10.1890/07-0043.1.
[24] R Core Team (2020). *Count Data And Overdispersion***.** R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/web/packages/GlmSimulatoR/vignettes/count_data_and_overdispersion.html

Model 1: Negative Binomial Model of Ridership including Zoning Regressors and Suburb Fixed Effect Dummy

The first model is designed to display the effect of only zoning regressors on ridership, paired with the suburb dummy to address all areas with 0 zoning values. This results in a 22 regressor model, in which the constant θ is a fitting parameter determined by the model to replicate the Poisson regression's fit:

$$\log(\theta(rides)) = \alpha + \beta_1 x_{B1} + \beta_2 x_{B2} + \beta_3 x_{B3} + \beta_4 x_{C1} + \beta_5 x_{C2} + \beta_6 x_{C3} + \beta_7 x_{DX} + \beta_8 x_{DC} + \beta_9 x_{DR} + \beta_{10} x_{DS} + \beta_{11} x_{M1} + \beta_{12} x_{M2} + \beta_{13} x_{M3} + \beta_{14} x_{PD} + \beta_{15} x_{PD} + \beta_{16} x_{PMD} + \beta_{17} x_{POS} + \beta_{18} x_{RM} + \beta_{19} x_{RS} + \beta_{20} x_{RT} + \beta_{21} x_T + \beta_{22} d_{suburb}$$

Figure 8: Negative Binomial Regression Equation, Model 1

Model 2: Negative Binomial Model of Ridership including Demographic Regressors and Suburb Fixed Effect Dummy

The second model includes only demographic regressors and the suburb fixed effect dummy, for a total of 13 regressors:

$$\log(\theta(rides)) = \alpha + \beta_1 x_{population} + \beta_2 x_{vehicles} + \beta_3 x_{employed} + \beta_4 x_{income} + \beta_5 x_{housing\ units} + \beta_6 x_{vacant\ units} + \beta_7 x_{bach.degree} + \beta_8 x_{unemployed} + \beta_9 x_{bus\ riders} + \beta_{10} x_{white} + \beta_{11} x_{black} + \beta_{12} x_{asian} + \beta_{13} d_{suburb}$$

Model 3: Negative Binomial Model of Ridership including Demographics, Zoning, Built Environment, and Suburb Fixed Effect Dummy

The final model utilizes a total of 35 covariates of interest with the same design as the prior models:

$$\log(\theta(rides)) = \alpha + \beta_1 x_{population} + \beta_2 x_{vehicles} + \beta_3 x_{employed} + \beta_4 x_{income} + \beta_5 x_{housing\ units} + \beta_6 x_{vacant\ units} + \beta_7 x_{bach.degree} + \beta_8 x_{unemployed} + \beta_9 x_{bus\ riders} + \beta_{10} x_{white} + \beta_{11} x_{black} + \beta_{12} x_{asian} + \beta_{14} x_{B1} + \beta_{15} x_{B2} + \beta_{16} x_{B3} + \beta_{17} x_{C1} + \beta_{18} x_{C2} + \beta_{19} x_{C3} + \beta_{20} x_{DX} + \beta_{21} x_{DC} + \beta_{22} x_{DR} + \beta_{23} x_{DS} + \beta_{24} x_{M1} + \beta_{25} x_{M2} + \beta_{26} x_{M3} + \beta_{27} x_{PD} + \beta_{28} x_{PD} + \beta_{29} x_{PMD} + \beta_{30} x_{POS} + \beta_{31} x_{RM} + \beta_{32} x_{RS} + \beta_{33} x_{RT} + \beta_{34} x_T + \beta_{35} d_{suburb}$$

Figure 9: Negative Binomial Regression Equation, Model 3

Figure 9 is replicated by the following call in R:

```
glm.nb(formula = rides ~ totpopulation + totvehicles + totemploy +  meanINCOME + tothousing +
totvacant + highered + totunemploy + busriders + totwhite + totblack + totasian + B1 + B2 + B3 + C1 +
C2 + C3 + DX + DR + DC + DS + M1 + M2 +  M3 + PD + PMD + POS + Tz + RS + RM + RT +
builtpost2000 + built70to99 + built40to69 + suburbf, data = finalset, init.theta = 4.501978028, link = log)
```

Figure 10: Negative Binomial Model with all regressors (Note: the model log-transforms the dependent 'rides' variable)

Results

Table 1: Block Group Level Predictors of Ridership, Summed at 0.5 Mile Radius from Stations

	Ridership 2016-2019		
	(Model 1)	*(Model 2)*	*(Model 3)*
totpopulation	-0.0000708***		-0.0000671**
	-0.0000264		-0.0000271
totvehicles	-0.0000154		-0.0000236
	-0.0000207		-0.0000212
totemploy	0.0000953***		0.0001112***
	-0.000035		-0.0000372
meanINCOME	0.0000036		0.000002
	-0.0000024		-0.0000026
tothousing	0.0000624*		-0.0000227
	-0.0000373		-0.0000447
totvacant	-0.0000815		-0.0000905
	-0.0000855		-0.0000961
highered	-0.0000645		-0.0000638
	-0.0000396		-0.0000399
totunemploy	0.0002156		0.0000005
	-0.0001352		-0.0001368
busriders	-0.0000854**		0.0000421
	-0.000036		-0.0000463
totwhite	0.0000249		0.0000430*
	-0.0000234		-0.0000254
totblack	-0.0000028		0.0000189
	-0.0000183		-0.000018
totasian	0.0000255		0.0000017
	-0.0000227		-0.0000233
B1		-0.0000001	-0.0000001
		-0.0000005	-0.0000005

	(1)	*(2)*	*(3)*
B2		-0.000001	-0.0000014
		-0.0000018	-0.0000017
B3		0.0000002	0.0000000
		-0.0000003	-0.0000003
C1		-0.0000011***	-0.0000008**
		-0.0000004	-0.0000004
C2		-0.0000003	0.0000004
		-0.0000008	-0.0000007
C3		-0.0000005	-0.0000004
		-0.0000025	-0.0000022
DX		-0.0000006**	-0.0000005
		-0.0000003	-0.0000003
DR		0.0000013	0.0000019
		-0.0000012	-0.0000013
DC		0.0000021	0.0000063*
		-0.0000035	-0.0000037
DS		-0.00000002	-0.0000004
		-0.0000005	-0.0000005
M1		-0.0000002	-0.00000002
		-0.0000002	-0.0000002
M2		-0.00000002	-0.00000002
		-0.0000001	-0.0000001
M3		-0.0000028	-0.0000024
		-0.0000033	-0.0000032
PD		0.0000001***	0.0000003***
		-0.00000004	-0.0000001
PMD		0.0000003	0.0000001
		-0.0000002	-0.0000002
POS		-0.0000007***	-0.0000002
		-0.0000002	-0.0000002

	(1)	(2)	(3)
T		-0.0000014	-0.0000056**
		-0.0000027	-0.0000028
RS		0.0000001*	0.0000001**
		-0.0000001	-0.0000001
RM		0.0000005**	-0.0000003
		-0.0000002	-0.0000003
RT		-0.0000001	0.00000002
		-0.0000001	-0.0000001
builtpost2000		0.0000972***	0.0000919***
		-0.0000227	-0.0000318
built70to99		-0.0000417	-0.0000127
		-0.0000282	-0.0000322
built40to69		0.0000345	0.0001118***
		-0.0000211	-0.0000338
suburbf1	-0.4944590***	-0.6819309***	-0.3537927
	-0.1719852	-0.2167021	-0.2402944
Constant	14.8229500***	14.9703800***	14.9444600***
	-0.2053683	-0.1728727	-0.240734
N	142	142	142
Log Likelihood	-2248.325	-2,243.901	-2,221.537
theta	3.1820080***	3.3985650***	4.5019780***
	(0.3595998)	(0.3858645)	(0.5157121)
Akaike Inf. Crit.	4524.649	4,536.368	4,517.074

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

All regression results should be interpreted as estimated percentage effects on ridership relative to a null '0' value. All zoning measurements are provided in square meter units such that the coefficients suggest the percentage effect of a one square meter increase in the zoning type in block groups half-mile proximate to stations.

Discussion and Interpretation of Results

Model Interpretation: Significant Indicators

1. Suburban Effect

While not significant in the full model (Table 1, column 3), the effect of the suburban dummy variable is notable for its high significance and negative coefficient in Models 1 and 2. The shift from significance – and a high percentage estimate of -49.4 and -68.2 in Models 1 and 2, respectively – to an insignificant, lower coefficient of -35.4 percent in Model 3 displays covariance between suburban stations and specific demographic variables common in suburbs. Interpreting the significant fixed effects of Models 1 and 2, location outside of the city's borders gives a mean 49.4% and 68.2% decrease in ridership, ceteris paribus to an urban station.

This relation, despite insignificance of the suburb dummy's estimate along with high standard error in Model 3, suggests unique traits of suburban areas that can be partially accounted for through other regressors regardless of the lack of Chicago zoning and thus zero values for all zoning covariates in these fringe areas of the system. In recognizing this effect, a simple scatterplot of log-transformed ridership as related to total employment (Figure 11), one of the only significant covariates in Model 3, displays clear linear relation, albeit with the caveat of a smaller sample size of suburban stations. Similar bivariate relations of interest could be inspected in the same manner, although employment's significance allows for cautious conclusions to be drawn:

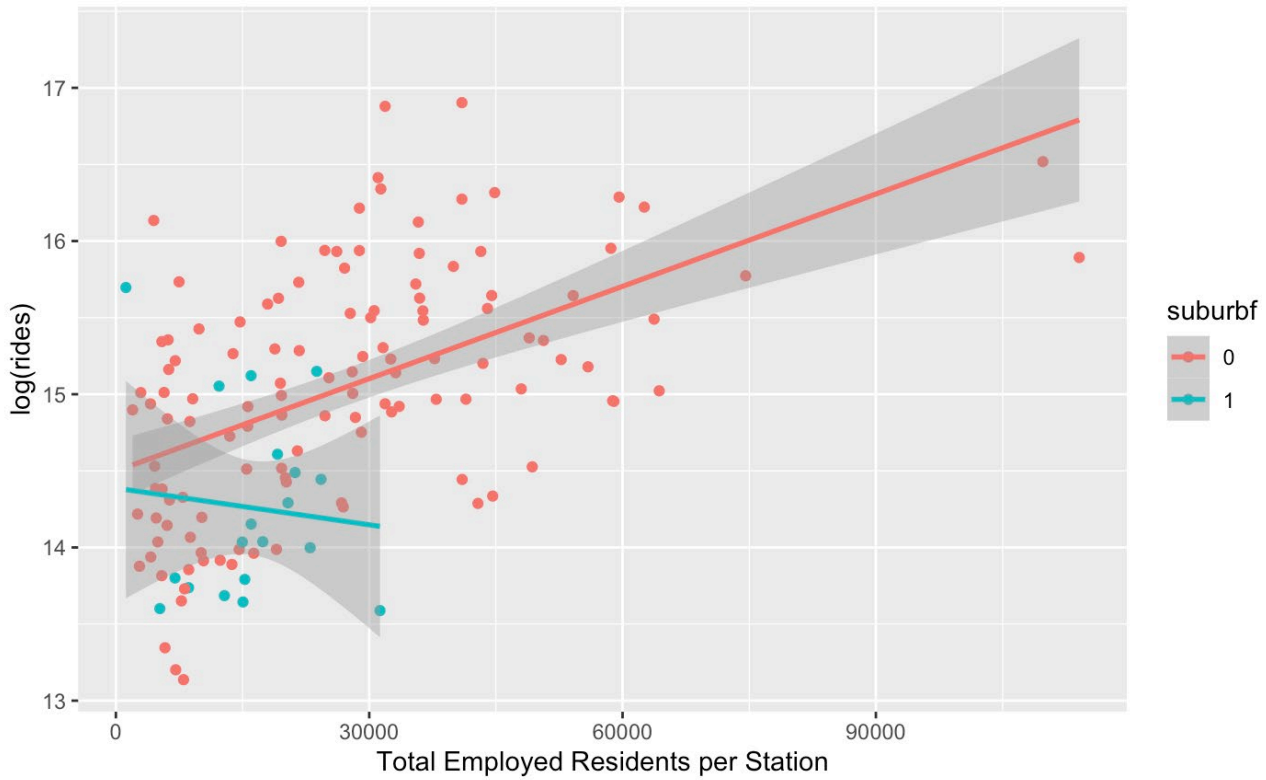

Figure 11: Employment v. logged ridership with bivariate regression lines

2. Employment

The high positive coefficient of .0001112 and high significance of the total employment variable suggests a ~0.0112% increase in ridership per employed resident over the period. This variable's significance in Model 3, while arguably biased upward given the massive percentage effect of a mean 37,820 employed residents per station area, is of interest – especially considering the bivariate relationship in Figure 11, in which suburban stations have an average 10,827 fewer employed than those within the city boundaries.

Addressing employment's role in the regression as a catch-all for larger working trends is inherently faulty, as the high presence of employed residents could be related to more working opportunities within the station's area and thus ridership from the station at the conclusion of the workday. This situation, while presumably common, is not accounted for in the CTA's data

collection methodology and thus warrants caution when interpreting a high coefficient such as employment per station area. Similarly, for stations with fewer surrounding block groups, such as those along the highway en route to O'Hare on the fringes of the Blue Line, summed counts are less than other stations and could bias the coefficient downwards in Models 1 and 3.

3. Zonings

Neighborhood Commercial District (C1)

The prevalence and relatively even dispersal of C1 districts along street corridors in the city appears to partially explain the coefficient of -0.0000008, which suggests a .00008% decrease in ridership for every square meter in the catchment area of each station. A mean of 154,620 square meters per station under the model's conditions would thus estimate a 12.37% decrease in ridership, the result of which should be questioned given a standard error of .0000004 or .00004%.

Of note in assessing this is the proximity of C1 zoning to low-ridership stations along the Green and Pink Lines' West Side extents as well as lower values due to not falling within centroids of buffer catchment areas on the higher ridership areas of the North Side's Red Line stretch. Along with this – and a common factor for all non-downtown (DC, DX, DR, and DS) zonings excluding Private Development (PD) – are the zero values present in the downtown area, which presumably make nearly all zoning types have lower coefficients given the outlying values at stations such as Clark and Lake, which sees the highest ridership in the period.

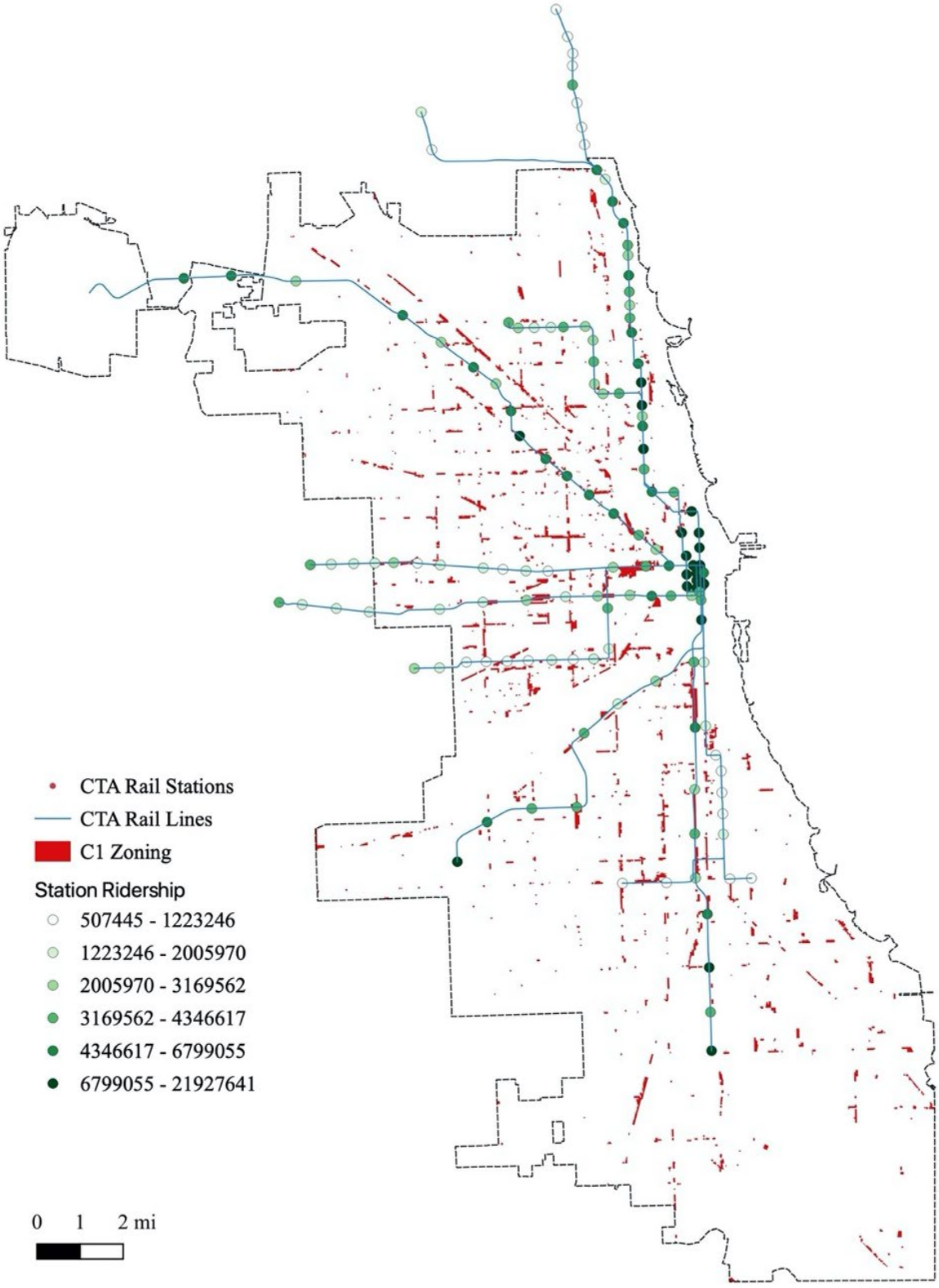

Figure 12: Neighborhood Commercial District (C1) Zoning. Source: Chicago Open Data Portal

Downtown Core District (DC)

The downtown core zoning, and the downtown zonings as a whole, are spatially constrained to the extent that their coefficients warrant wary approaches. This is both due to upper quantile ridership stations within the catchment area, many of which are outliers, and the lack of the zoning anywhere else in the city to provide evidence of validity as an estimator without considering the economic and entertainment opportunities within downtown, approximately a square mile of which is almost completely covered by DC zoning (Figure 12).

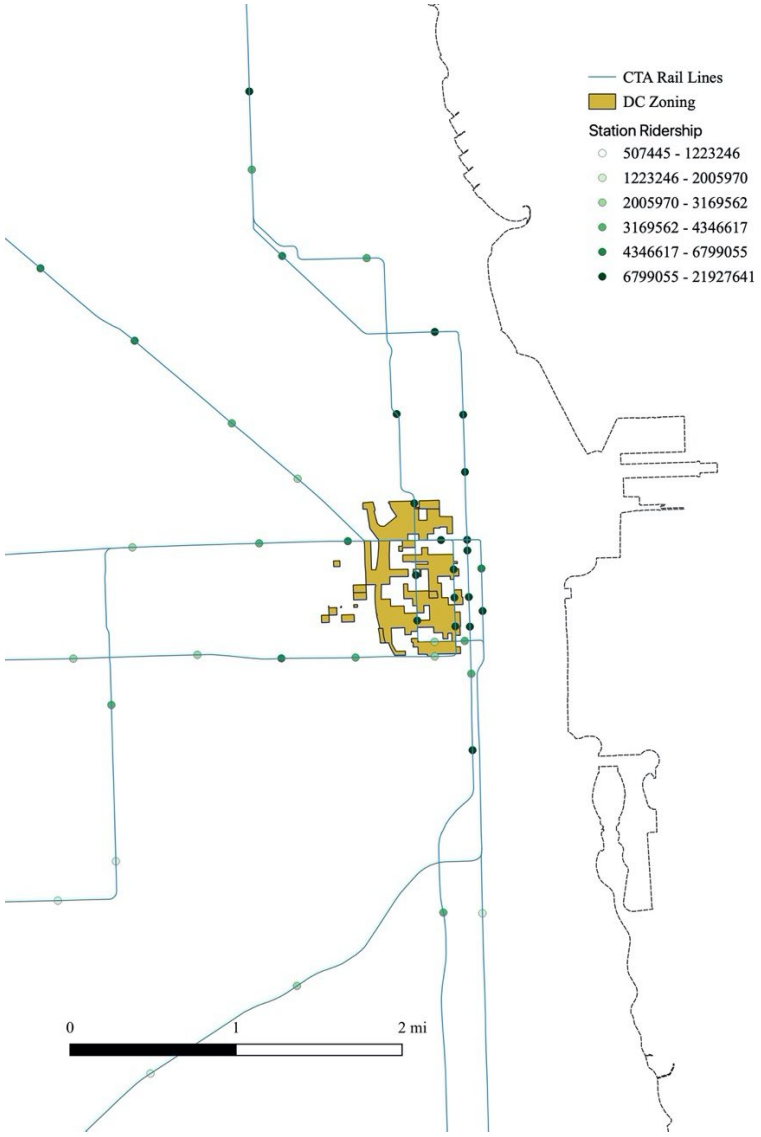

Figure 13: Downtown Core (DC) Zoning. Source: Chicago Open Data Portal

Planned Developments (PD)

The vagary of the Planned Development zoning arises from the nature of the zoning type as city approved development by a variety of interests, which accounts for areas ranging from Soldier Field and Navy Pier to Midway and O'Hare, with a multitude of smaller developments accounting for the sparse presence across the city's commonly residential extents (Figure 13). As these are often interspersed with other surrounding zoning types solely due to their distinct character as singular, city-approved developments, the variation in land use around them and thus also within the catchment areas, e.g., the 'Loop' area of downtown, where zoning appears to alternate between PD (Figure 14) and DC zonings (Figure 13).

Given the unique nature of the PD designation, associating the positive coefficient of .0000001, or .00001% increase per square meter (albeit with an extremely high estimate-equivalent standard error) with effect on ridership is difficult. As further discussed in the 'Limitations' section of the study, assigning value to planned development zoning as a transit estimator would be better served through a parcel-by-parcel reassignment to other zoning types or the creation of new types – such as academic and entertainment land uses – to grant a degree of granularity to the model. Therefore, this zoning type can cautiously be interpreted at a macro-scale in its proximity to and association with places which attract transit riders and are actively planned around transit opportunities. In this interpretation, the positive coefficient is reasonably associated with a myriad of opportunities associated with these areas throughout the city.

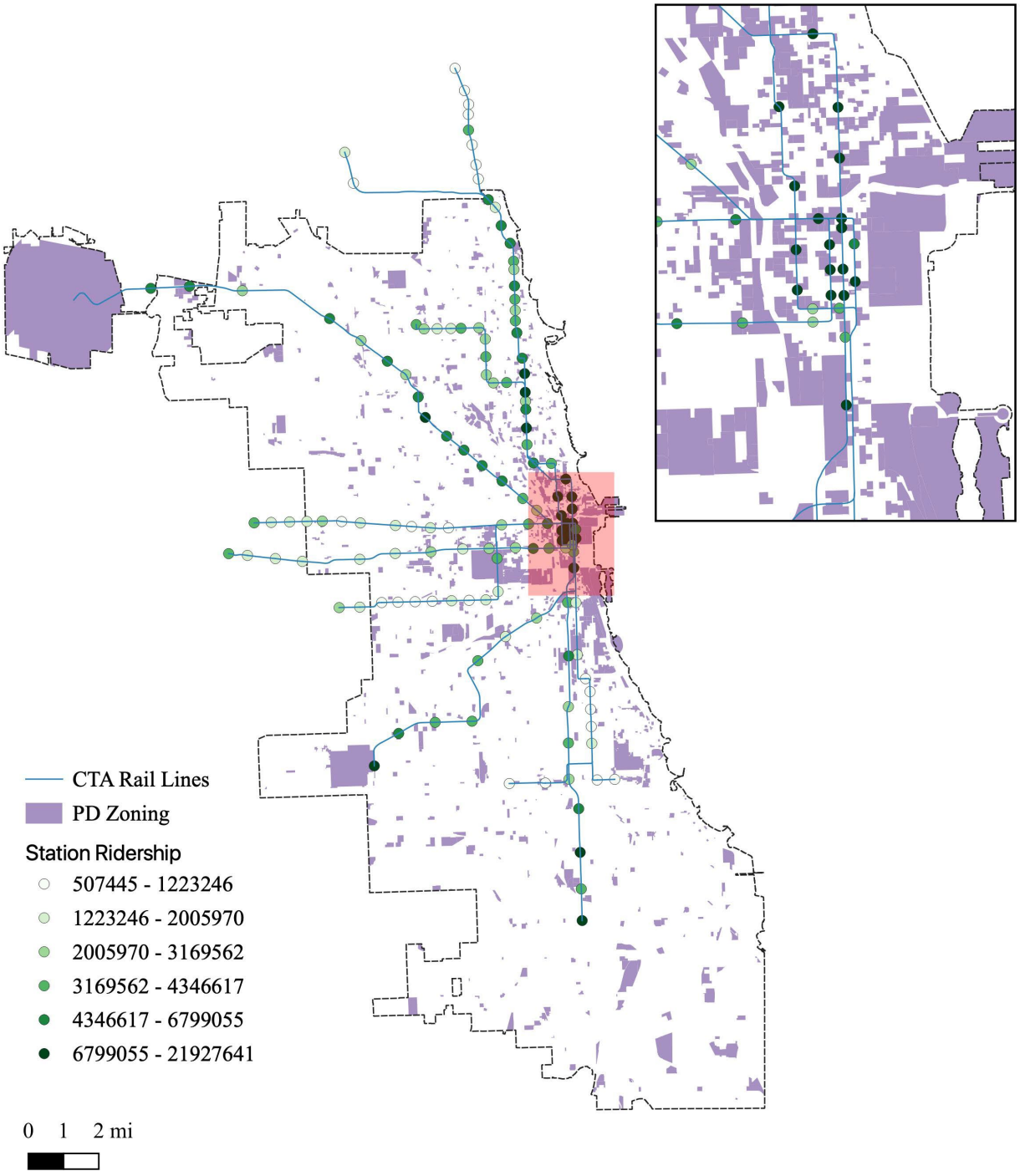

Figure 14: Planned Development (PD) Zoning. Source: Chicago Open Data Portal

Residential Single Unit District (RS)

The Residential Single Unit District is the most common in the city by area, existing in nearly all regions of the city whose residences are not characterized by densely built environments and thus RM or DR zonings (Note the absence of RS zoning in the commonly industrial West Side of the city along the Pink, Green, and Blue Line in Figure 15). Given its prevalence in the city and high areal coverage in nearly all stations with non-zero values, RS zoning is relatively uniform in its high values along the non-central stretches of multiple lines. Much like other zoning types, minor distinctions in this type – past nuanced ordinances among the aforementioned subtypes – would greatly alleviate cautions toward assigning any causal legitimacy to the estimate. Given this fact combined with the high standard error, which could cause a zero estimate at its low-bound, this estimate is unfortunately minimal in relevant conclusions despite clear its clear predictive power in the model.

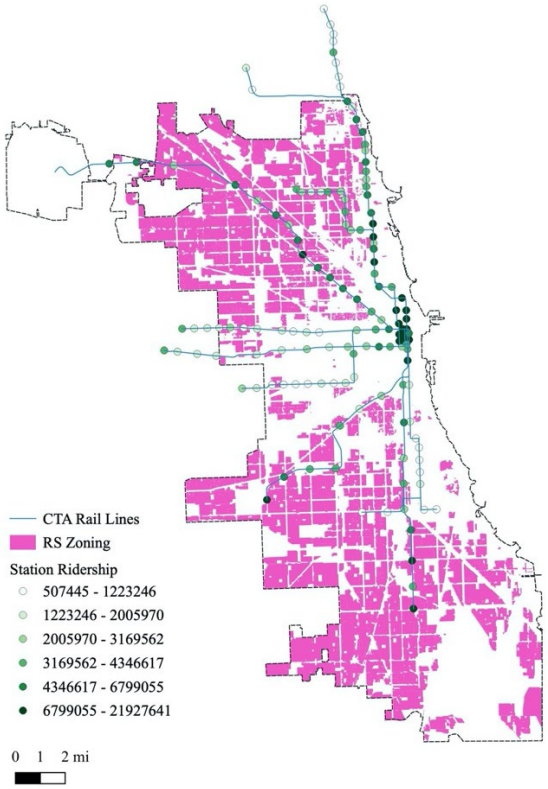

Figure 15: Residential Single Unit District (RS) Zoning. Source: Chicago Open Data Portal

Transit District (T)

Transit zoning occupies the smallest area of the significant zoning regressors, and only exists at non-zero values within the catchment areas of 14 stations. As displayed in Figure 16, all relevant areas to the analysis excepting one small district located North of the 35th Street Red and Green Line stations occur on the North and Northwest sides of the city.

With a coefficient of -.0000056, the associated .00056% decrease in ridership per square meter is presumably related to the built environment of these areas. As stated in Figure 2, this designation is relatively vague in its use as a preservationist measure; for example, the east to west stretch along the Blue Line simply designates the area of the 606, a former elevated rail track turned recreational trail. As a valid means of commuting for local workers in a lateral direction not served by the CTA's rail offerings, a zoning region such as this is interpreted by the model as negative, a fact at least partially attributable to omitted variable bias. Others have less immediately clear use; the aforementioned area on 35th street is located along the highway with no discernible function outside of storage or as reserve for future development. With the constraint of the city's intent with these spaces being highly varied, interpretation of Transportation Zoning is generally hindered at the system-wide scale.

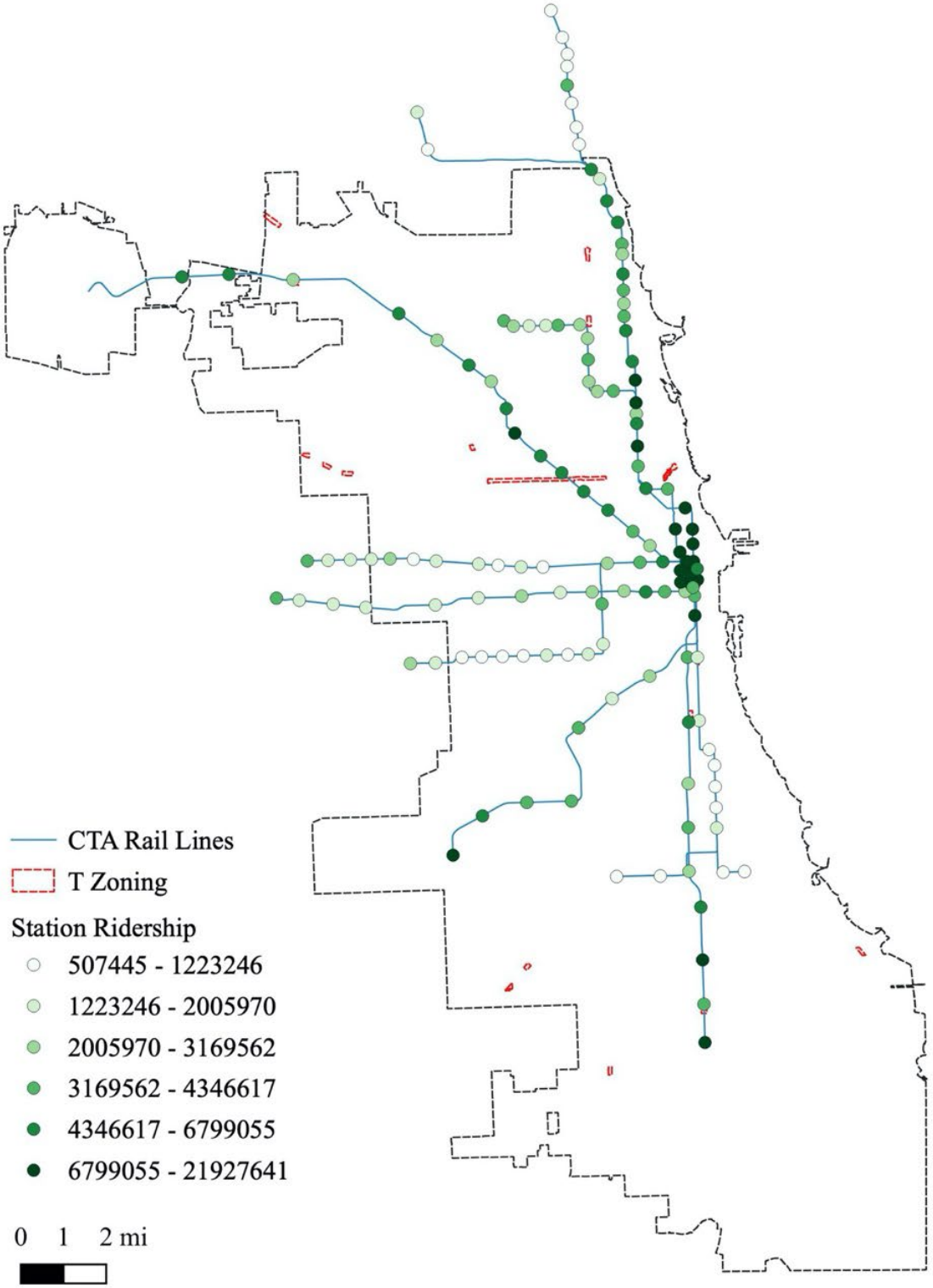
CTA Rail Lines
T Zoning
Station Ridership
507445 - 1223246
1223246 - 2005970
2005970 - 3169562
3169562 - 4346617
4346617 - 6799055
6799055 - 21927641
0 1 2 mi

Figure 16: Transportation (T) Zoning. Source: Chicago Open Data Portal

4. Built Dates

1940-1969

Chicago's development history is defined by eras to an extent that it is easily recognized in and visualized through the station summations (Figure 17). The highly significant positive coefficient of .0001118, or an estimated .01% increase in ridership per structure, appears to be most associated with clustering around high use stations along the Red Line's northern extent, which sees high rates of this variable throughout the Gold Coast and closer to downtown within Streeterville, as well as in Wicker Park and Logan Square along the midsection of the Blue Line. Development analysis of this kind would be best served by smaller-scale approaches, but the clear patterns nonetheless make interesting suggestions as to ridership and development correlation in this era.

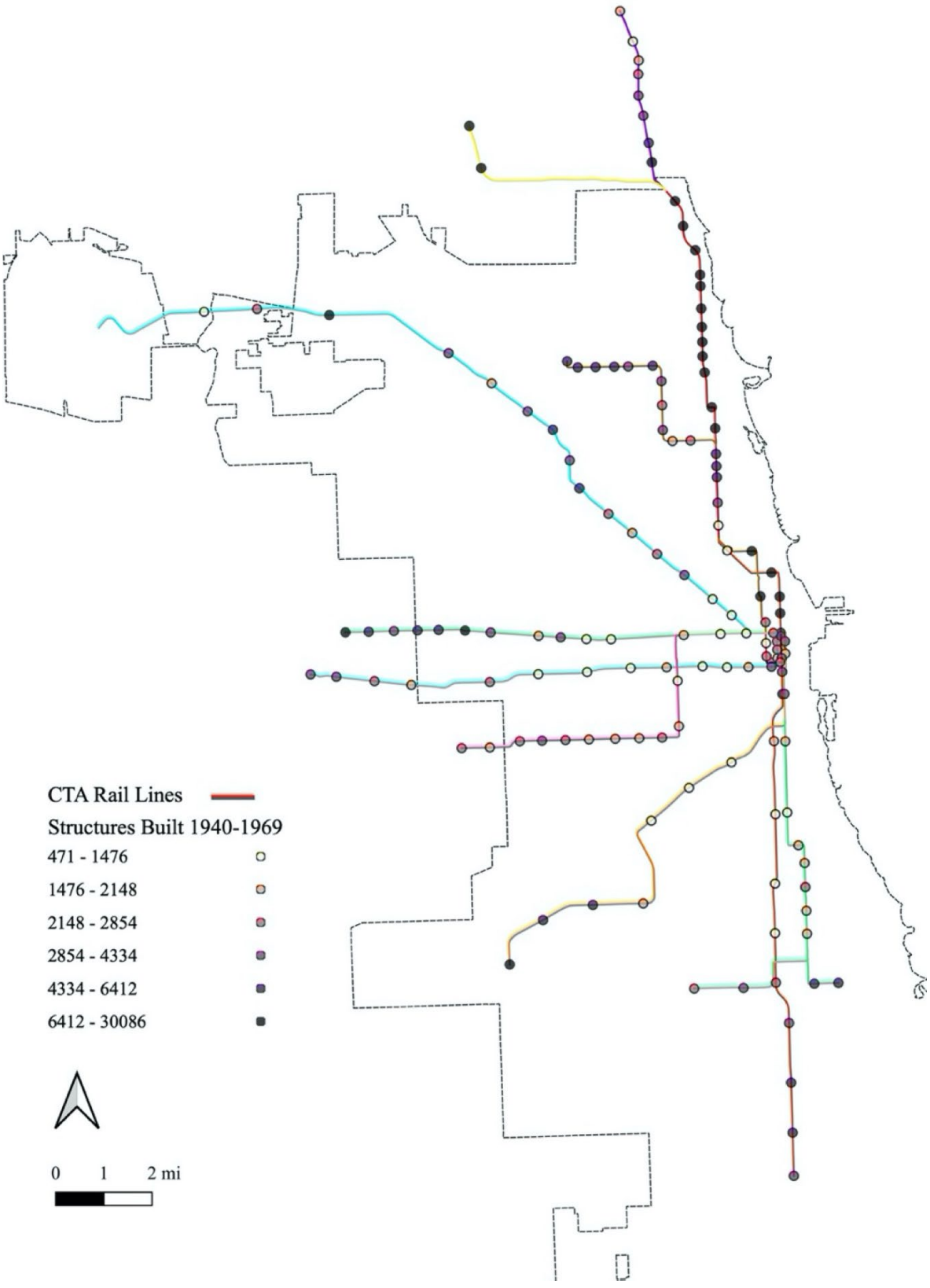

Figure 17: Structures Built 1940-1969, Source: NHGIS

Post-2000

The prevalence of upper quantile values among stations in the high ridership downtown and Red Line displays clear linkage between recent development in revitalized, high investment corridors of the city. As such, the highly significant coefficient of .0000919 is reasonably high given proximity to areas with high tourism, entertainment, and economic activity alike.

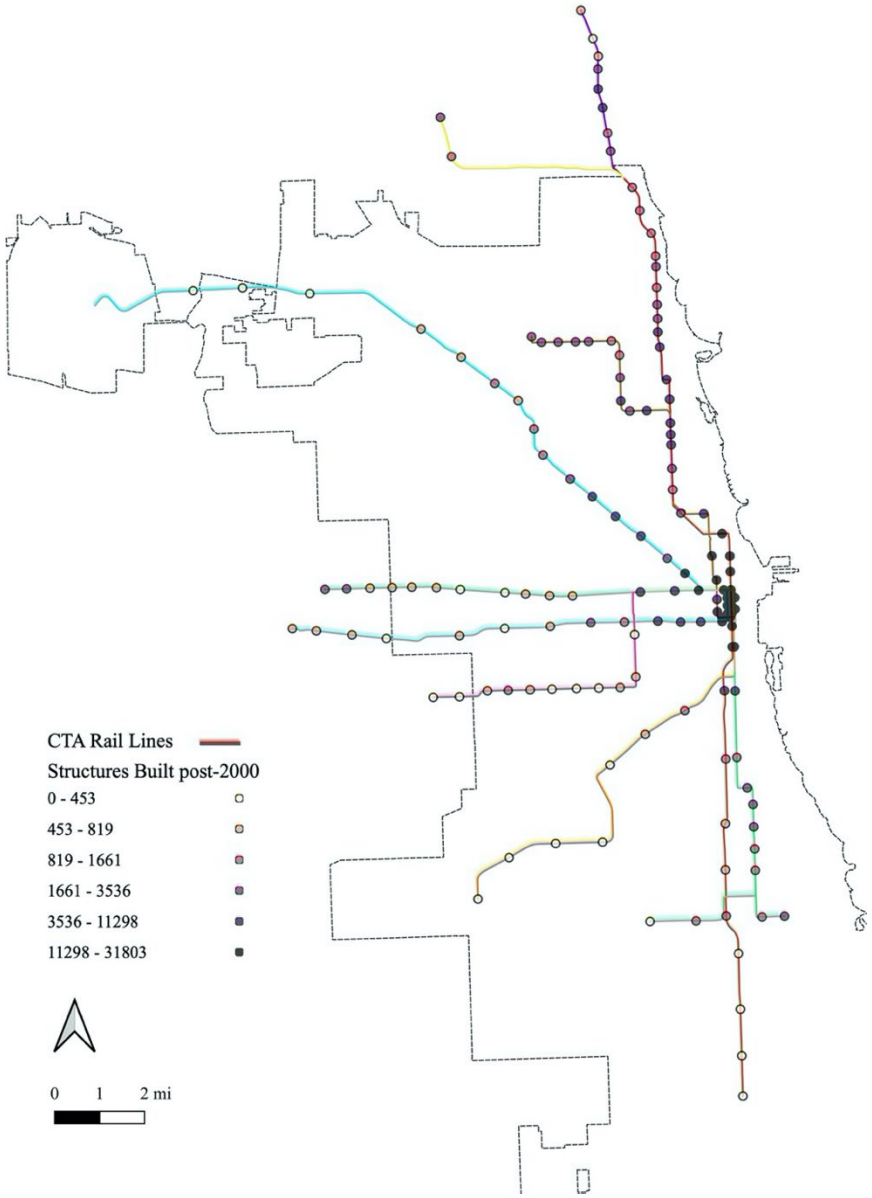

Figure 18: Structures Built Post-2000. Source: NHGIS

Concluding Thoughts

The creation of a quantitative model inherently fails to account for the countless urban processes shaping transit ridership during this era, a fact shown both in the constraints of making causal claims and in the statistical form of high standard errors and few significant estimates. The fine tuning of the model given omitted variable biases is of concern, but overfitting to an extent that the model underperforms, especially given relatively high covariance between certain indicators, could prove difficult in a revised regression setting. However, clear patterns of transit use emerge among significant variables such that recognition of demographic, zoning, and built environment relations to ridership in the 2016-2019 era appear to be apt in estimating current and future effects to be reckoned with as the CTA continues to adapt and maintain efficiency.

Limitations

Chicago-Specific Zonings and Suburban Disparities

Use of Chicago's zoning data as a proxy for land use causes the issue of proper grouping of zone types created through more than a century of changing practices and ordinances. The minimal, often vague variations between different zonings allows for possible misrepresentation of the zoning data aggregated within certain block group areas, especially given the spatial variations present across Chicago. Consideration various subtypes as part of single zoning types was chosen for ease of regression and resulting analysis, but with this less granular data comes lesser recognition of unique zoning effects. Among the broadest zoning types is that of planned developments, denoted by the city as "PD-(X)" with a number following based upon the date of ordinance/introduction. Since the zoning data used for this study included more than 1200 unique "PD" areas, their variety of uses could not be well accounted for without manual zone-by-zone reassignment to other categories or the creation of new subgroups. With the time constraints and

focus of the project, PD zones were grouped together in the interest of demonstrating the effect of privately held zoning area on transit use.

Unfortunately, the incorporation of zoning limited the assessment of stations to those which fall within the city's boundaries. This excludes all suburban stations from assessment by land use variables, an unfortunate condition given these stations' consistently low unstandardized ridership metrics compared to the rest of the system. This could be addressed by use of satellite data or a combined zoning index across all of Cook County, yet none currently exists in the public space. The creation of a dummy for suburban block groups was created as a fixed effect for the lack of zoning, such that all zoned block groups are marked "0." While the suburb dummy was highly useful in determining a highly significant lower use for stations in the suburbs, I recognize that using a fixed effect of suburban location fails to address the many land use disparities and transit preferences present in these areas. These are only partially accounted for in the variables discussed above, which is immediately shown in the maximum ~.5 R^2 statistic. Confounding factors and possible variables in the continuation of this study would incorporate Metra ridership for commuters, a factor which is not only outside of the current data but also generally untethered to the walkability assumptions in this study due to the lesser density of stations and car-centric suburban design.

Regional Block Group Densities

While the half-mile radius was chosen to best represent transit-use patterns, the assignment of block groups as proximate to the station is not without faults. This is primarily clear in the disparate counts of block groups within the radius of each station, which varies significantly in different regions of the city, most clearly in the less dense extremities of the system (Figure 2). Given the differing counts of nearby block groups, the model's aggregation

and thus relative equivalent importance per station is biased towards an understanding of stations nearer to the mean number of surrounding block groups such that stations with only one nearby block group see its statistics overrepresented.

Faulty and Missing Values in Data

Unfortunately, the O'Hare station block group had missing data in the ACS to the extent that it would have severely biased the data downwards to assign mean or "0" values to the data. This led to the removal of the station from the regression despite recognition of its importance as a travel node and a station at the end of a line. The negative effect of the station's removal is partially tempered by the PD zoning attribute of the whole O'Hare area, which leaves little variety and would only increase the sample size of a single zoning variable. Similarly, the singular airport block group creates a small sample size for summation to the station.

Data Granularity due to CTA System Design

As discussed earlier, the CTA's data provides a high level of day-to-day detail yet encounters vagaries surrounding individual rider details and trip plans. The aggregation of ridership, especially at transit nodes such as the Loop area, causes difficulties for quantifying ridership by line due to the option for riders to take multiple trains after paying the fare at that station. Therefore, relating node stations to others existing on multiple lines grants only minimal insight due to a significant disparity in the stations' functions and options.

Further Studies and Next Steps

Considering the wealth of data available from the CTA and the ever-increasing quality of publicly available data, I believe that a similar study could be completed with varying methods based on data choice, hopefully to more statistically and socially significant findings.

While the variables identified as transit related in this study and the literature preceding it are of constant interest, they can also be presumed to constantly change in their importance given shifts in the built environment, cultural norms and preferences, and the spatial distribution of populations and workers. This study was designed with recency in mind yet chose not to address the COVID-19 pandemic in the interest of maintaining relative uniformity in variable effects without introducing the discontinuities resulting from city and system-wide lockdowns in March 2020. I hope that the study's observed trends apply to post-pandemic transit use and economic realities but recognize that individual transit preferences resulting from the pandemic could be misaligned with any representations gained through analysis of the 2016-2019 era.

Limitations of processing power and time constraints made the use of time-series data considerably more difficult but given a focus on day-to-day or month-to-month variables instead of rolling counts, a viable natural experiment could be conducted with similar regressors. This methodology could provide causal conclusions to underuse, albeit with strenuous data collection and possible constraints to smaller, case-study level scales of stations with confounding results in non-causal, estimate based analyses such as this study.

Bibliography

Brown, Anne E., et al. A Taste for Transit? Analyzing Public Transit Use Trends among Youth. Journal of Public Transportation, 19 (1): 49-67. 2016.

Cervero, R, Radisch, C. Travel Choices in Pedestrian Versus Automobile Oriented Neighborhoods. *UC Berkeley: University of California Transportation Center*. 1995. https://escholarship.org/uc/item/7cn9m1qz

Erick Guerra, Robert Cervero, and Daniel Tischler. 2012. "The Half-Mile Circle: Does It Best Represent Transit Station Catchments?" Transportation Research Record: Journal of the Transportation Research Board, 2276: 101–109.

Graehler, Michael & Mucci, Alex & Erhardt, Gregory. *Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?. 2012.*

Hertz, Daniel Kay. "Opinion: The Green Line's Waiting Game." *South Side Weekly*. 14 Nov. 2017. https://southsideweekly.com/opinion-waiting-game-green-line-cta/.

Hilbe, Joseph M. *Negative Binomial Regression*. Cambridge: Cambridge University Press, 2007. doi:10.1017/CBO9780511811852.

Litman, T. Evaluating public transit benefits and costs: Best Practices Guidebook, Victoria Transport Policy Institute. 2021.

Marshall Lindsey, Joseph L. Schofer, Pablo Durango-Cohen, Kimberly A. Gray, Relationship between proximity to transit and ridership for journey-to-work trips in Chicago. Transportation Research Part A: Policy and Practice, Volume 44, Issue 9, Pages 697-709. 2010.

Min, B., Lee, G., Kim, S. Effects of land-use characteristics on transport mode choices by purpose of travel in Seoul, South Korea, based on spatial regression analysis. Sustainability (Switzerland), 13 (4), art. no. 1767, pp. 1-22. 2021.

National Museum of American History Behring Center, Smithsonian Institute. "Chicago, the

Transit Metropolis," https://americanhistory.si.edu/america-on-the-move/essays/chicago-transit-metropolis.

R Core Team (2020). *Count Data And Overdispersion***.** R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://cran.rproject.org/web/packages/GlmSimulatoR/vignettes/count_data_and_overdispersion.html

Sabyasachee Mishra, Timothy F. Welch, Manoj K. Jha. Performance indicators for public transit connectivity in multi-modal transportation networks, Transportation Research Part A: Policy and Practice, Volume 46, Issue 7, Pages 1066-1085. 2012.

Staff. "CTA Celebrates 25th Anniversary of Orange Line Service to SW Chicago & Midway Airport." *CTA.* CTA. 30 Oct. 2018. https://www.transitchicago.com/orange25/.

Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 16.0. Minneapolis, MN: IPUMS. 2021. http://doi.org/10.18128/D050.V16.0

"TNC Use, Transit, and Vehicle Ownership in Chicago." *Center for Neighborhood Technology*. 2019. https://www.cnt.org/blog/tnc-use-transit-and-vehicle-ownership-in-chicago.

"Transit-Oriented Development in the Chicago Region: Efficient and Resilient Communities for the 21st Century." *Center for Neighborhood Technology*. 2013.

Ver Hoef, Jay & Boveng, Peter. (2007). *Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data?.* Ecology. 88. 2766-72. 10.1890/07-0043.1.

Zotti, Ed. "The Case for Rail Transit Expansion in the Chicago Central Area" National University Rail Center – NURail. 2016.